1-1-2013

RNA and DNA Sequence Analysis of the Human Transcriptome

Jonathan M. Toung
University of Pennsylvania, jmtoung@gmail.com

RNA and DNA Sequence Analysis of the Human Transcriptome

Abstract
The manifestation of phenotype at the cellular and organismal level is determined in large part by gene expression, or the transcription of DNA into RNA. As such, the study of the transcriptome, or the characterization and quantification of all RNA produced in the cell, is important. Recent advances in sequencing technology have allowed for unprecedented interrogation of the transcriptome at single-nucleotide resolution. In the first part of this thesis, we use RNA-Sequencing (RNA-Seq) to study the human B-cell transcriptome and determine the experimental parameters necessary for sequencing-based studies of gene expression. We discover that deep sequencing is necessary to detect fully and quantify accurately the complexity of human transcriptomes. Furthermore, we find that at high sequencing depths, the vast majority of transcribed elements in human B-cells are detected.

In the second part of this thesis, we utilize the sequence information provided by RNA-Seq to analyze systematic differences between DNA and RNA sequence. The transmission of information from DNA to RNA is a critical process and is expected to occur in a one-to-one fashion. By comparing the DNA sequence to RNA sequence of the same individuals, we found all 12 types of RNA-DNA sequence differences (RDDs), the majority of which cannot be explained by known mechanisms such as RNA editing or transcriptional infidelity. We developed computational methods to robustly identify RDDs and control for false positives resulting from genotyping, sequencing, and alignment error. Finally, we explore the genetic basis of RDD levels, or the proportion of reads at a site bearing the sequence difference allele. In particular, we analyzed the levels of RNA editing in unrelated and related individuals and found that a significant portion of individual variation in A-to-G editing levels contains a genetic component.

In summary, our results demonstrate that RNA-Seq is a powerful technique for comprehensive and quantitative analysis of gene expression. In addition, the resolution offered by RNA-Seq enables a detailed view of sequence differences between RNA and DNA. Future work will focus on understanding the mechanisms and factors influencing RDDs. Our results suggest that RDD levels may be considered a quantitative and heritable phenotype; as such, a genetic approach may be a sensible method for finding the determinants and mechanism of RDDs.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

First Advisor
Frederic Bushman

Second Advisor
Nancy Zhang

Keywords
Computational biology, Genome analysis, Next-generation sequencing, RNA Editing, RNA-Sequencing, Transcriptome analysis

Subject Categories
Bioinformatics | Genetics

RNA AND DNA SEQUENCE ANALYSIS OF THE HUMAN TRANSCRIPTOME

Jonathan M. Toung

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania in
Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2013

Frederic Bushman Professor of Microbiology Supervisor of Dissertation	Nancy Zhang Professor of Statistics Co-supervisor of Dissertation

Maja Bucan
Professor of Genetics
Graduate Group Chairperson

Dissertation Committee:
Christopher D. Brown, Assistant Professor of Genetics
Gregory R. Grant, Research Assistant Professor of Genetics
Shane T. Jensen, Associate Professor of Statistics (chair)
Thomas A. Jongens, Associate Professor of Genetics
Ponzy Lu, Professor of Chemistry

Dedication

For Esther and my family.

Acknowledgements

While many a times it has felt that the pursuit of this degree has been a solitary and singular effort, I would be foolish to not acknowledge the many people that have helped me tremendously along the way.

This journey began by Dr. Ponzy Lu intermittently badgering me to pursue a Ph.D since the first time I sought out his advice as a freshman in the fall of 2006. My undergraduate mentor, thesis committee member, and life coach – I am truly indebted to your tireless support of me.

I thank Dr. Vivian Cheung for her support of my studies. I thank Michael Morley for teaching me all there is to know about bioinformatics and programming. Much of what I have accomplished today computationally would not have been possible without your help. Thank you Dr. Isabel Xiaorong Wang. I truly appreciate your willingness to sit me down in your office and help with my committee meetings, my papers, my public talks. I thank past and present members of the Cheung lab for their help and support: Wendy Ankener, Will Bernal, Lauren Brady, Alan Bruzel, Jim Devlin, Susannah Elwyn, Brittany Gregory, Anna Lee, Sen Sen Liu, Yaojuan Liu, Colleen McGarry, Allison Richards, and Elizabeth So.

I would not have been able to complete this degree had it not been for individuals who saw a precarious situation and did everything in their power to rectify it. Dr. Rick Bushman, I thank you so much for your support and encouragement throughout the year. Thank you for being the first to provide a space for me in your lab to finish my studies. You have gone above and beyond your duties as an advisor to vouch and fight for me. Dr. Nancy Zhang, I thank you for your support as my advisor. I cannot forget your efforts

throughout the situation to make sure I could progress and move on with my studies. Your statistical insights into the project have always been inspiring, and I am truly grateful to have learned so much from you.

Dr. John Hogenesch, you have been a tremendous help for my last paper. Your encouragement and advice on how to deal with the situation was invaluable. Dr. Greg Grant, chapter 4 would not have been possible without you. Thank you for your advice and devotion to the project.

To members of my committee, Dr. Greg Grant, Dr. Shane T. Jensen, Dr. Tom Jongens, Dr. Ponzy Lu, Dr. Nancy Zhang and members of my academic review committee, Dr. Maja Bucan, Dr. Rick Bushman, and Dr. John Hogenesch – thank you all for getting me through this wild ride.

Thank you Dr. Warren Ewens and Dr. Mingyao Li for your help and guidance on my studies.

Thank you to everyone in the Genomics and Computational Biology program. A special thanks to Hannah Chervitz and all that you do to make our lives easier as students.

To my parents, thank you for your support and lessons to never give up. I'm glad we made it through another journey.

And lastly, to my dear girlfriend Esther. You deserve this degree as much as I do. You have put up with a lot in the past three years. It has been an emotional roller coaster and you have been with me every step of the way. Here's to better days ahead.

Someone once told me the wise saying that life is never as good as you imagine it to be nor as bad as you think it is. It will be just okay.

ABSTRACT

RNA AND DNA-SEQUENCE ANALYSIS
OF THE HUMAN TRANSCRIPTOME

Jonathan M. Toung

Frederic Bushman

Nancy Zhang

The manifestation of phenotype at the cellular and organismal level is determined in large part by gene expression, or the transcription of DNA into RNA. As such, the study of the transcriptome, or the characterization and quantification of all RNA produced in the cell, is important. Recent advances in sequencing technology have allowed for unprecedented interrogation of the transcriptome at single-nucleotide resolution. In the first part of this thesis, we use RNA-Sequencing (RNA-Seq) to study the human B-cell transcriptome and determine the experimental parameters necessary for sequencing-based studies of gene expression. We discover that deep sequencing is necessary to detect fully and quantify accurately the complexity of human transcriptomes. Furthermore, we find that at high sequencing depths, the vast majority of transcribed elements in human B-cells are detected.

In the second part of this thesis, we utilize the sequence information provided by RNA-Seq to analyze systematic differences between DNA and RNA sequence. The transmission of information from DNA to RNA is a critical process and is expected to occur in a one-to-one fashion. By comparing the DNA sequence to RNA sequence of the

same individuals, we found all 12 types of RNA-DNA sequence differences (RDDs), the majority of which cannot be explained by known mechanisms such as RNA editing or transcriptional infidelity. We developed computational methods to robustly identify RDDs and control for false positives resulting from genotyping, sequencing, and alignment error. Finally, we explore the genetic basis of RDD levels, or the proportion of reads at a site bearing the sequence difference allele. In particular, we analyzed the levels of RNA editing in unrelated and related individuals and found that a significant portion of individual variation in A-to-G editing levels contains a genetic component.

In summary, our results demonstrate that RNA-Seq is a powerful technique for comprehensive and quantitative analysis of gene expression. In addition, the resolution offered by RNA-Seq enables a detailed view of sequence differences between RNA and DNA. Future work will focus on understanding the mechanisms and factors influencing RDDs. Our results suggest that RDD levels may be considered a quantitative and heritable phenotype; as such, a genetic approach may be a sensible method for finding the determinants and mechanism of RDDs.

Table of Contents

List of Tables

Chapter 5. Genetic Basis of RNA-DNA Sequence Differences

List of Figures

Chapter 2. RNA-Sequence Analysis of Human B-Cells

Chapter 3. RNA-DNA Sequence Differences in Humans

Chapter 4. Detection Theory in Identification of RNA-DNA Sequence Differences Using RNA-Sequencing

Chapter 5. Genetic Basis of RNA-DNA Sequence Differences

Chapter 1. Introduction

1.1 The study of the transcriptome

The realization of phenotype at the cellular and organismal level is determined in large part by a more proximal phenotype – gene expression, or the transcription of DNA into RNA. As such, the study of the transcriptome, or the characterization and quantification of all RNA produced in the cell at a given time or under a particular condition, is critical for a deeper understanding of all biological pathways and processes. Over the past few decades, the role of complexity at the RNA level in contributing to disease manifestation and phenotypic variation has become increasingly more apparent (Cooper et al. 2009; Licatalosi & Darnell 2010). From the point of transcription, RNA is subject to a wide range of processes such as alternative transcription initiation (Davuluri et al. 2008; Moore & Proudfoot 2009), alternative polyadenylation (Di Giammartino et al. 2011; Lutz 2008), alternative splicing (Cooper et al. 2009; Maniatis & Tasic 2002), RNA editing (Gott & Emeson 2000; Knoop 2011), and other post-transcriptional modification events (Karijolich & Yu 2011; G. Martin & Keller 2007; Rottman et al. 1994). In recent years, methodological advances in sequencing and bioinformatics have enabled genome-wide analyses of RNA at unprecedented levels of resolution and depth, allowing for comprehensive profiles of RNA species and variation (Djebali et al. 2012; Nagalakshmi et al. 2008). These developments lend insight into the contribution of RNA to overall biological diversity and cellular function.

In this thesis, we study the human transcriptome using RNA-Sequencing (RNA-Seq) technology. In particular, we quantify the expression levels of genes and their transcripts and determine the appropriate experimental parameters for sequencing-based

studies. Next, we examine systematic RNA-DNA sequence differences (RDDs) – discrepancies between DNA and RNA that may result from RNA editing, transcriptional infidelity, or other unknown mechanisms. Lastly, we analyze variation in levels of RDD, or the percentage of transcripts at a particular site that are altered, across unrelated individuals and assess the degree to which this individual variation is determined by genetic factors. The studies performed and methodologies used in this thesis demonstrate the coming of a new age in modern biology where challenges lie not in the procurement of but rather the analysis of data.

1.2 Hybridization- and sequencing-based methods for transcriptome analysis

Traditionally, hybridization-based methods such as microarrays were used to study gene expression and quantify RNA in a global manner (Schena et al. 1995). The main strategy for hybridization-based approaches involves incubating fluorescently labeled complementary DNA (cDNA) fragments with microarray chips that are fixed with oligonucleotide probes containing known target sequences. Gene expression levels are subsequently quantified by fluorescent detection of the probe-target pairs. Studies using microarrays have provided key insights into the genetics and regulation of gene expression (Brem et al. 2002; Morley et al. 2004), cancer biology (Bittner et al. 2000; Golub et al. 1999), and various biological processes such as cell growth (Iyer et al. 1999) and the cell cycle (Cho et al. 1998; Cho et al. 2001). Nonetheless, array methodologies are subject to limitations that include background noise resulting from cross-hybridization (Okoniewski & Miller 2006), dependence on known targets, and small dynamic range. Some specialized microarrays have been designed, however, to address some of these issues, such as the use of genomic tiling microarrays to detect unknown transcripts (Bertone et al. 2004).

To overcome the disadvantages of hybridization-based approaches, several sequencing-based methods have been developed to study the transcriptome. In contrast to microarrays, sequencing-based approaches directly determine the sequence of the underlying sample. Initially, Sanger sequencing of cDNA or EST libraries was used to determine gene sequences (Boguski et al. 1994), although this method is relatively expensive and low-throughput. Tag-based approaches were developed to overcome these

limitations and generate digital gene expression profiles in a high-throughput manner. These tag-based methods include serial analysis of gene expression (SAGE) (Velculescu et al. 1995) cap analysis of gene expression (CAGE) (Kodzius et al. 2006), and massively parallel signature sequencing (MPSS) (Brenner et al. 2000). In general, these approaches involve the introduction of recognition sites to the ends of cDNA and cutting by restriction endonucleases to create tags that can be isolated and cloned for high-throughput sequencing (Harbers & Carninci 2005). These approaches, however, are expensive as they rely on Sanger sequencing, and their output is restricted to certain regions of the transcripts, limiting the analysis of alternatively-spliced isoforms. Recently, a high-throughput sequencing approach called RNA-Sequencing (RNA-Seq) has been developed that overcomes many of the shortcomings of hybridization-based methods and preexisting sequencing-based approaches.

1.3 Next-generation sequencing methods for transcriptome analysis

Apart from providing the scientific community with a reference genome, the Human Genome Project spurred advancements in sequencing technology. Termed 'next-generation sequencing technology', these high-throughput sequencing platforms have become more readily accessible to researchers within the past decade, replacing the more expensive capillary sequencing methods and dramatically decreasing the cost of DNA sequencing (Mardis 2008; Metzker 2010; Shendure & Ji 2008). The ability of next-generation sequencing methods to provide an unprecedented amount of sequence information at a relatively low cost has enabled both whole-genome sequencing and *de novo* assembly of novel genomes (R. LiW. Fan, et al. 2010; Locke et al. 2011) in addition to whole-genome re-sequencing of organisms with reference genomes to catalogue and annotate genetic variants (1000 Genomes Project Consortium 2010; Bentley et al. 2008; J. I. Kim et al. 2009; Korbel et al. 2007; S. C. Schuster et al. 2010; J. Wang et al. 2008; Wheeler et al. 2008).

In addition to the direction application of next-generation sequencing to sequencing of DNA (DNA-Seq), a whole host of sequencing-based assays have been developed using next-generation sequencing platforms to interrogate genome-wide profiles of mRNA (Mortazavi et al. 2008; Nagalakshmi et al. 2008), RNA secondary structure (Kertesz et al. 2010; Underwood et al. 2010), transcription factor binding (X. Chen et al. 2008; Farnham 2009; Johnson et al. 2007; Wederell et al. 2008), chromatin states and histone modifications (Barski et al. 2007; Mikkelsen et al. 2007; Schones et al. 2008), DNase hypersensitivity (Boyle et al. 2008), and DNA methylation status (Cokus et al. 2008; Meissner et al. 2008). In particular, high-throughput sequencing of RNA or RNA-

Seq has revealed insights into the complexity of the transcriptome, uncovering new classes of small RNA (Lau et al. 2006; Malone & Hannon 2009; Taft et al. 2010), discovering novel transcripts and gene fusions (Bruno et al. 2010; Levin et al. 2009; Lu et al. 2010; Maher et al. 2009), and expanding the catalogue of alternatively-spliced transcripts (Pan et al. 2008; E. T. Wang et al. 2008).

As of this writing, the most commonly used next-generation sequencing platforms are the Illumina Genome Analyzer (Bentley et al. 2008) and, more recently, the Illumina HiSeq machines, the Roche 454 Genome Sequencer FLX machines (Margulies et al. 2005; Rothberg & Leamon 2008), and the SOLiD system from Life Technologies (McKernan et al. 2009). Other next-generation sequencing manufacturers or platforms include Complete Genomics (Drmanac et al. 2010), Helicos BioSciences (Braslavsky et al. 2003; Ozsolak et al. 2009), Ion Torrent from Life Technologies (Rothberg et al. 2011), and Pacific Biosciences (Eid et al. 2009). For general applications, Illumina is currently the market leader in the next-generation sequencing space because of low error rates and high yield (Luo et al. 2012; Minoche et al. 2011; Quail et al. 2012), although the various platforms have advantages that are better suited for different experiments. For example, the Roche 454 system delivers long reads necessary for many metagenomics studies (Turnbaugh et al. 2009; Wommack et al. 2008), Helicos sequencers allow for direct sequencing of RNA without conversion to cDNA (Ozsolak & Milos 2011; Ozsolak et al. 2009), and the Pacific Biosciences platform allows for real-time monitoring of the sequencing reaction (Metzker 2009) and detection of modified nucleotides (Flusberg et al. 2010). While the exact methodology for each system differs, in general, the next-

generation sequencing workflow involves library or template preparation, sequencing and imaging, and alignment and assembly.

1.3.1 Library preparation and sequencing

The main steps of library preparation consist of isolation of nucleic acid material, shearing of the sample into smaller fragments, addition of adapter sequences to allow for PCR amplification, and immobilization of the fragments to a surface. Illumina uses a technique called solid-phase amplification (Fedurco et al. 2006) whereby fragments hybridize to primers covalently attached to a glass slide or 'flow cell'. Fragments are clonally amplified using isothermal 'bridging' amplification, resulting in high-density 'clusters' originating from the same template (Bentley et al. 2008). In contrast, Roche 454 and SOLiD by Life Technologies employ a different technique called emulsion PCR (Dressman et al. 2003) which involves bead capture of fragments and immobilization of the beads through chemical crosslinking (Kim et al. 2007) or deposition into wells (Leamon et al. 2003).

The next steps in the workflow calls for sequencing and imaging of the clonally amplified templates. Illumina utilizes a strategy termed 'cyclic reversible termination' which involves repeated cycles of nucleotide incorporation, fluorescence imaging, and cleavage (Bentley et al. 2008; Metzker 2005). The methodology begins with the incorporation of a modified nucleotide containing a reversible blocking group that terminates DNA synthesis after addition of a single base. The identity of the incorporated base and an associated quality score reflecting the probability of a misidentification is determined through fluorescence imaging, followed by a cleavage step that removes the inhibitor group from the DNA chain. Roche 454 uses a method called 'pyrosequencing'

that measures the release of inorganic pyrophosphate to infer the incorporation of nucleotides (Ronaghi et al. 1996; Ronaghi et al. 1998). The pyrosequencing method does not use modified nucleotides to terminate DNA synthesis, but instead relies on the addition of nucleotides in limiting amounts and recording the luminescence following release of pyrophosphate to infer the underlying DNA sequence (Margulies et al. 2005). Lastly, the SOLiD system by Life Technologies uses a 'sequencing by ligation' approach that uses DNA ligase (Landegren et al. 1988) and base-encoded probes (Shendure et al. 2005). Briefly, fluorescently labeled probes are hybridized to the complementary sequencing template. DNA ligase is then added, non-ligated probes are washed away, and the identity of the incorporated probe along with a quality score reflecting the probability of an incorrect call is then determined by imaging.

1.3.2 Genome assembly and alignment

For genome studies, the first step in the processing of sequence reads obtained from next-generation sequencing platforms is assembly and alignment (Flicek & Birney 2009). Assembly of sequencing reads is necessary in cases where a reference genome does not exist. Prior to the advent of next-generation sequencing data, Sanger sequencing technology provided long reads (~800 nucleotides or nt), and assembly algorithms resolved the overlap between these long reads. For the short read (36 to ~300 nt) output of next-generation sequencing platforms, new computational methods have been developed to address the assembly of fragments of small lengths (Zhang et al. 2011). Two main types of strategies exist for *de novo* assembly: string-based methods, implemented by a Greedy-extension algorithm, and graph-based models such as the overlap-layout-consensus method and the de Bruijn graph approach. Programs that use

string-based methods, such as QSRA (Bryant et al. 2009), VCAKE (Jeck et al. 2007), SHARCGS (Dohm et al. 2007), and SSAKE (Warren et al. 2007) are more appropriate for small genomes with short reads, whereas software that utilize overlap-layout-consensus, such as Forge (Diguistini et al. 2009), Shorty (Hossain et al. 2009), CABOG (Miller et al. 2008), and Edena (Hernandez et al. 2008) are more suited for small genomes with long reads (Zhang et al. 2011). For larger genomes, programs that use the de Bruijn graph method (Idury & Waterman 1995; Pevzner et al. 1989) are more fitting; algorithms that use this approach include ALLPATHS-LG (Gnerre et al. 2011), SOAPdenovo (R. Li, H. Zhu, et al. 2010), ABySS (Simpson et al. 2009), Euler-USR (Chaisson et al. 2009), Velvet (Zerbino & Birney 2008), ALLPATHS (Butler et al. 2008), and Euler-SR (Chaisson & Pevzner 2008).

In cases where a reference genome exists, alignment is necessary to determine the most likely origin for each sequencing read. Early methods relied on the use of hash-table data structures to index sequence data and rapidly search for alignments. One strategy, used by programs such as Eland (proprietary software of Illumina), RMAP (A. D. Smith et al. 2008), MAQ (H. Li et al. 2008), ZOOM (H. Lin et al. 2008), SeqMap (Jiang & Wong 2008), CloudBurst (Schatz 2009), and SHRiMP (Rumble et al. 2009), involves hashing the short reads and subsequently scanning through the reference genome for alignments. These software programs are relatively efficient in terms of memory usage, but they require a lookup of the entire reference genome regardless of the number of alignments in the dataset. The other methodology utilized by hash-based programs is to hash the reference genome instead. Programs that use this approach include BFAST (Homer et al. 2009), GSNAP (T. D. Wu & Nacu 2010), MOM (Eaves & Gao 2009),

PASS (Campagna et al. 2009), ProbeMatch (Y. J. Kim et al. 2009), SOAPv1 (R. Li et al. 2008), and Stampy (Lunter & Goodson 2011). These methods use a constant amount of memory for a given reference genome, which may be large depending on the size and (Lunter & Goodson 2011) complexity of the genomic sequence. A third type of strategy for alignment of next-generation sequencing data indexes the reference genome using the Burrows-Wheeler Transform (BWT), a data compression algorithm that allows for memory-efficient storage and fast string matching (Burrows & Wheeler 1994). Programs that use this approach include Bowtie (Langmead et al. 2009), BWA (H. Li & Durbin 2009), and SOAPv2 (R. Li, C. Yu, et al. 2009). While BWT-based aligners are fast and memory-efficient, they tend to be less accurate and sensitive than hash-based algorithms (Grant et al. 2011; Lunter & Goodson 2011).

1.3.3 Transcriptome assembly and alignment

For transcriptome studies, the assembly and alignment challenge is complicated by unequal coverage resulting from gene expression differences, strand-specific expression of transcripts, and alternative splicing. While the reference genome can serve as a scaffold in genome alignment for many model organisms, the existence of a truly comprehensive transcriptome database that accounts for all possible gene isoforms does not exist for most organisms. Thus, even for studies involving humans where a reference genome is available, the ability of assemblers to discover gene isoforms *de novo* is important. Transcriptome assembly programs mainly follow either a reference-based strategy or *de novo* approach, or some combination of the two (J. A. Martin & Wang 2011).

The *de novo* approach for transcriptome assembly involves searching for overlapping sequence between short reads and assembling them using a de Bruijn approach (Pevzner et al. 2001) to form transcripts. Programs using this strategy include Multiple-k (Surget-Groba & Montoya-Burgos 2010), Rnnotator (J. Martin et al. 2010), Trans-ABySS (Robertson et al. 2010), and Trinity (Grabherr et al. 2011). *De novo* assembly for organisms with smaller genomes has been successful, although for more complicated eukaryotic organisms, intense computing resources are necessary. Some disadvantages of *de novo* methods include the dependence on high sequencing depths (J. Martin et al. 2010) and sensitivity to sequencing error and chimeric transcripts (Cocquet et al. 2006).

Another strategy for transcriptome analyses of next-generation sequencing data uses the reference genome as a basis for guided assembly. The general strategy first involves alignment of the short reads to the genome using a splice-aware aligner, or aligner that allows for the introduction of gaps in locations of spliced introns. Examples of splice-aware aligners include BLAT (Kent 2002), GSNAP (T. D. Wu & Nacu 2010), MapSplice (K. Wang et al. 2010), QPALMA (De Bona et al. 2008), SOAPv2 (R. Li, C. Yu, et al. 2009), RUM (Grant et al. 2011), SpliceMap (Au et al. 2010), and TopHat (Trapnell et al. 2009). These various aligners use different approaches for the alignment of spliced reads. Early algorithms performed alignment to a transcriptome database consisting of known isoform transcripts supplemented with sequences surrounding potential exon-exon junctions (E. T. Wang et al. 2008); however, this method fails to detect unannotated exons and unconventional splicing events. The strategy used by Tophat involves a sequential approach: first, reads are aligned to the genome, second,

exons are inferred from regions of the genome with coverage, and lastly, reads that failed to align to the genome previously are mapped to junctions created between the proposed exons (Trapnell et al. 2009). This method performs well only for highly abundant isoforms, as the exons of low-expressing transcripts may not be properly identified. Another shortcoming of this strategy is that alignments to the genome take precedence over spliced alignments, which is problematic for alignments to homologous sequences in the genome such as processed pseudogenes. The methodology used by RUM addresses this issue, as RUM seeks and compares alignments to the genome and transcriptome (Grant et al. 2011). The approach taken by GSNAP searches for alignments to known junctions and also attempts to predict novel splicing events through a probabilistic model dependent on donor and acceptor sites in the surrounding genomic sequence (T. D. Wu & Nacu 2010). Comparative and benchmark analyses have been performed on these various splice-aware aligners, and the results indicate that GSNAP and RUM have the highest accuracy in terms of overall alignment and splice junction detection (Grant et al. 2011).

Following alignment of the reads to the genome with a splice-aware aligner, reads that overlap are clustered to build a graph representing all possible splice variants, and finally, isoforms are resolved by traversing through the graph. Programs that adopt the reference-based approach to transcriptome assembly include Cufflinks (Trapnell et al. 2010), G-Mo.R-Se (Denoeud et al. 2008), and Scripture (Guttman et al. 2010). Cufflinks assembles gene isoforms by first creating an overlap graph from all reads within a locus in the genome and then creates transcripts by determining the minimum number of traversals through the graph that explains the splice junctions observed (Trapnell et al. 2010). In contrast, Scripture forms a graph for each chromosome, inserting edges

between bases connected by a splice junction, and then assembles transcripts by searching for traversals through the graph that are supported by statistically significant expression (Guttman et al. 2010). The differences in graph formation and traversal methods between Cufflinks and Scripture suggests that Cufflinks may be more conservative in isoform detection, and a recent comparison between Cufflinks and Scripture has shown that indeed Cufflinks has higher specificity and sensitivity for finding known introns (Robertson et al. 2010).

A third approach to transcriptome assembly involves a combination of *de novo* and reference-based assembly. Algorithms that use both approaches combine the ability of *de novo* methods to find novel isoforms and the sensitivity and specificity in splice junction detection of referenced-based strategies. Some programs, such as Trans-ABySS (Robertson et al. 2010) and Oases (Schulz et al. 2012), begin by a referenced-based assembly of reads, followed by *de novo* assembly of reads that failed to align to the genome. For situations where a high quality reference genome does not exist, an alternative approach involves *de novo* assembly of reads into contigs and subsequent alignment of the contigs to the reference genome. The reference genome, and in some cases protein sequences, can be used to join and fill in gaps between contigs to form longer transcripts (Crawford et al. 2010; Surget-Groba & Montoya-Burgos 2010).

1.3.4 Variant and genotype calling

The goals of genome sequencing projects include the comprehensive characterization of genetic variation in a population (1000 Genomes Project Consortium 2010; Ju et al. 2011; Mills et al. 2011), the analysis of genetic adaptation or molecular evolution (Turner et al. 2010; Xia et al. 2009; Yi et al. 2010), and the mapping of disease

loci (Ng, Bigham, et al. 2010; Ng, Buckingham, et al. 2010; Ng et al. 2009; O'Roak et al. 2011; Tennessen et al. 2012). These studies, among other applications of high-throughput sequencing, rely critically on the accurate identification of genetic variants and genotyping of individual samples using next-generation sequencing data. Many frameworks and computational pipelines for variant and genotype calling have been developed. In general, they involve many steps of calibration, filtering, and statistical analyses. The general approach starts with methods for reducing base-calling errors in raw reads along with further recalibration of base quality scores, local realignment, and filtering of duplicate reads after assembly and alignment of reads. Variants, or sites where at least one of the observed bases is different from the reference, are then identified and the corresponding genotypes are inferred. For studies involving multiple samples, datasets can be pooled across individuals to increase the power of detection. Smaller variants, such as single-nulceotide polymorphisms (SNPs) and short insertions and deletions, require different algorithms and strategies than large structural variants such as insertions, deletions, inversions, duplications, copy-number variations (CNVs), and translocations. As such, many different programs aimed at different variants have been developed for the comprehensive analyses of DNA-Seq datasets.

The proper identification of variants and determination of genotype is affected by sequencing and alignment error. Thus, various precautionary steps are taken to reduce the amount of false positives resulting from spurious signals in next-generation sequencing datasets. The first step in most frameworks for variant and genotype calling involves optimization of base calling procedures. After assembly and alignment of reads, further measures, such as quality score readjustment, local realignment, and filtering of reads

with low mapping scores, are taken. Base-calling algorithms generally output a base call in addition to a Phred quality score that represents the probability of an error in base calling (Ewing & Green 1998). Reports have shown, however, that these quality scores co-vary with sequencing platform, reaction cycle, and local sequence features (Brockman et al. 2008; M. Li et al. 2004; R. Li, Y. Li, et al. 2009; Nakamura et al. 2011). As such, various de-noising algorithms, such as Ibis (Kircher et al. 2009) and BayesCall (Kao & Song 2011; Kao et al. 2009) for Illumina, Pyrobayes (Quinlan et al. 2008) for Roche 454, Rsolid (H. Wu et al. 2010) for SOLiD, and SysCall for Illumina and SOLiD (Meacham et al. 2011) have been developed to reduce the error rates for base calling. After assembly and alignment, many variant calling pipelines remove reads aligning to multiple locations in the genome or reads with low mapping quality scores. In order to account for the correlation of sequencing error with the position of a base within the read (Balzer et al. 2010; Minoche et al. 2011), some variant calling algorithms further recalibrate sequencing base quality scores (DePristo et al. 2011; H. Li et al. 2009; R. Li, Y. Li, et al. 2009; R. Li, C. Yu, et al. 2009; McKenna et al. 2010).

Many algorithms have been developed for the identification and genotyping of SNPs. Early methods for genotyping SNPs relied on filtering out bases with Phred quality scores less than 20 and identifying a site as heterozygous if the percentage of non-reference bases at any given site is between 20 and 80% (Harismendy et al. 2009; J. Wang et al. 2008). This method is a good approximation, however, only when the sequencing depth is large. Furthermore, this simple cut-off approach does not take into account the quality scores of individual bases and it does not provide an uncertainty measure for the genotype inferred. Thus, several probabilistic models that take into

consideration various parameters such as allele frequencies and patterns of linkage disequilibrium have been developed to address these issues (Nielsen et al. 2011). These algorithms include GATK (DePristo et al. 2011; McKenna et al. 2010), MAQ (H. Li et al. 2008), QCall (Le & Durbin 2011), Samtools (H. Li et al. 2009), SOAPv2 (R. Li, C. Yu, et al. 2009), and VarScan (Koboldt et al. 2009). After genotyping has been performed on individual samples, datasets can be pooled to filter out singletons and assemble a list of SNP variants.

In addition to SNP discovery programs, many different approaches have been developed for the identification of short insertions and deletions (indels) and larger structural variants. These methods include Dindel (Albers et al. 2011), inGAP (Qi et al. 2010), GATK (DePristo et al. 2011; McKenna et al. 2010), MoDIL (S. Lee et al. 2009), Samtools (H. Li et al. 2009), and VarScan (Koboldt et al. 2009) among others (Krawitz et al. 2010; D. R. Smith et al. 2008). For the detection of larger structural variants, algorithms rely critically on the use of paired-end reads, or sequencing of fragments from two ends of the library read, to infer deviations from expected distances or relative orientations between two locations in the genome (Campbell et al. 2008; Kidd et al. 2008; Korbel et al. 2007). This paired-end mapping technique has allowed for genome-wide characterization of large structural variants such as insertions, deletions, inversions, duplications, copy-number variations (CNVs), and translocations. Computational methods developed for robust detection of such variants include ABI SOLiD Software Tools (McKernan et al. 2009), BreakDancer (K. Chen et al. 2009), MoDIL (S. Lee et al. 2009), PEMer (Korbel et al. 2009), Pindel (Ye et al. 2009), SegSeq (Chiang et al. 2009), and Variation Hunter (Hormozdiari et al. 2009).

1.3.5 Gene expression profiling

Applications of next-generation sequencing technology to the study of RNA have led to many insights into the complexity of the transcriptome. In particular, RNA-Seq has been used to probe various aspects of the transcriptome such as gene expression (Graveley et al. 2011; Mortazavi et al. 2008; Nagalakshmi et al. 2008), alternative splicing (Cloonan et al. 2008; Sultan et al. 2008; Wilhelm et al. 2008), alternative transcription start site usage (Trapnell et al. 2010; Twine et al. 2011), and alternative polyadenylation (Ozsolak et al. 2010; Sandberg et al. 2008; E. T. Wang et al. 2008) among others. These features of the transcriptome, along with many others, are critically dependent on accurate methods for quantifying gene expression from RNA-Seq data. Reported initially in 2008, the standard measurement of gene expression is 'FPKM', or fragments per kilobase of exon model per million mapped reads (Mortazavi et al. 2008). This calculation of gene expression accounts for differences in input sizes across experiments through normalization of the total number of reads aligned and corrects for variation in gene size by dividing by the sum of exon lengths. Improving upon the standard 'FPKM' measurement, various other adjustments have been proposed to address non-uniform coverage of gene transcripts (Roberts et al. 2011) and uncertainty resulting from reads that align to multiple locations in the genome (B. Li et al. 2010). For studies of differential gene expression, different computational pipelines such as Cufflinks (Trapnell et al. 2012), DEGseq (L. Wang et al. 2010), DESeq (Anders & Huber 2010), DSS (H. Wu et al. 2012), edgeR (Robinson et al. 2010), GFOLD (Feng et al. 2012), and Myrna (Langmead et al. 2010) have been designed to normalize measurements across samples and accurately find gene transcripts that change in expression levels.

1.4 RNA-DNA sequence differences

The single-nucleotide resolution provided by next-generation sequencing technology permits detailed examination of DNA and RNA sequence. In particular, the interrogation of the genome and transcriptome by deep sequencing permits systematic comparisons between DNA and RNA sequence. In general, the transmission of sequence information from DNA to RNA is expected to occur in a one-to-one fashion (Crick 1970). There are, however, known exceptions to this rule: transcriptional infidelity, or errors introduced by RNA polymerase during the synthesis of RNA, and RNA editing, the modification of RNA transcripts by various mechanisms to alter the original RNA sequence specified in the genome. Recently, with advances in next-generation sequencing technology, many researchers have performed genome-wide studies comparing DNA and RNA sequences from the same sample and found many more editing events than previously known in addition to RNA-DNA sequence differences (RDDs) that cannot be explained by known mechanisms.

1.4.1 Transcriptional infidelity

Errors introduced during transcription by RNA polymerases are rare, with an estimated frequency of less than 1 in 10^5 in bacteria and eukaryotes (Blank et al. 1986; de Mercoyrol et al. 1992; Rosenberger & Hilton 1983). Strategies for ensuring high transcriptional accuracy include discriminative substrate selection (Kireeva et al. 2008; Libby & Gallant 1991; Temiakov et al. 2004; Vassylyev et al. 2007) and proofreading capabilities (Borukhov et al. 1993; Erie et al. 1993; Izban & Luse 1992; Kuhn et al. 2007; Sydow & Cramer 2009). Despite these mechanisms to ensure faithful transcription, errors do occur, albeit at a low frequency. Misincorporation events can result from mispairing

of the incoming nucleotide with the template (Sydow et al. 2009), non-templated incorporation from an abasic template site (Damsma et al. 2007), transcription of DNA lesions (Brueckner et al. 2007; Damsma et al. 2007), or template misalignment (Kashkina et al. 2006; Pomerantz et al. 2006). Furthermore, studies on RNA polymerase show that misincorporation rates are correlated with the type of mismatch (Sydow et al. 2009) and surrounding sequence context (Kashkina et al. 2006).

1.4.2 RNA editing

RNA editing refers to targeted sequence alteration of the RNA transcript, resulting in sequence that is different from the underlying DNA template. Originally discovered in trypanosomes as an insertion event of four uridine residues to restore a reading-frame shift in the DNA (Benne et al. 1986), the term has since come to encompass a diverse set of sequence revisions of the RNA transcript in a wide range of organisms (Gott & Emeson 2000; Knoop 2011). In lower organisms, various types of editing mechanisms have been uncovered since the initial discovery of insertional editing in trypanosomes. For example, paramyxoviruses control the expression of two different isoforms of phosphoprotein by introducing either one or two additional guanosines in the mRNA transcript co-transcriptionally (Cattaneo et al. 1989; S. M. Thomas et al. 1988; Vidal et al. 1990a). The insertion event is always found to occur within a short stretch of guanosines, and thus a stuttering mechanism has been proposed in which the viral polymerase repeatedly copies a cytosine in the template (Vidal et al. 1990b). In myxomycetes, various types of co-transcriptional editing events have been discovered, such as C-to-U substitutions and insertions of adenosines, cytidines, uridines, and dinucleotides (Gott et al. 1993; Hendrickson & Silliker 2010; Horton & Landweber 2000,

2002; Mahendran et al. 1991). Other examples of RNA editing in lower organisms include pyrimidine and purine transitions, transversions of guanosine to cytidine, and conversions of uridine to purines among other sequence revisions in dinoflagellates (Dorrell & Howe 2012; Jackson et al. 2007; S. Lin et al. 2002; Zauner et al. 2004) and U-to-C conversions in placazoa (Burger et al. 2009).

In plants, RNA editing occurs in the form of pyrimidine exchanges in mitochondrial and chloroplast transcripts. Since the initial discovery of C-to-U substitutions to reconstitute evolutionarily conserved amino acids in wheat and evening primrose mitochondria (Covello & Gray 1989; Gualberto et al. 1989; Hiesel et al. 1989), other instances of C-to-U editing have been observed in the mitochondria of many different land plants (Fang et al. 2012; Hiesel et al. 1994; Malek et al. 1996; Sper-Whitis et al. 1996; Sper-Whitis et al. 1994; W. Wang et al. 2012). Less common than the commonly observed C-to-U editing, U-to-C exchanges have also been observed in the mitochondria hornworts (Steinhauser et al. 1999; Yoshinaga et al. 1996), lycophytes (Grewe et al. 2009), and some seed plants (Gualberto et al. 1990; W. Schuster et al. 1990). Shortly after the discovery of mitochondrial RNA editing, C-to-U editing of chloroplasts was reported in maize (Hoch et al. 1991) and subsequently in all land plants with the exception of the marchantiid liverworts (Freyer et al. 1997; Inada et al. 2004; Kugita et al. 2003; Tillich et al. 2005; Tsudzuki et al. 2001; Wolf et al. 2004). Studies have demonstrated that chloroplast editing in plants is essential for protein function (Bock et al. 1994) and that specificity is conferred by local sequences (Bock et al. 1996; Chaudhuri et al. 1995; Sutton et al. 1995). Recent efforts to find the mechanism underlying plant organelle editing have implicated the RNA-binding pentatricopeptide

repeat (PPR) protein family as playing a key and necessary role (S. R. Kim et al. 2009; Kotera et al. 2005; Okuda et al. 2010; Robbins et al. 2009; Yu et al. 2009; Zehrmann et al. 2009; Zhou et al. 2009).

In metazoa, two main types of editing are known to exist: A-to-I changes result from deamination of adenosine to inosine by ADAR, a family of adenosine deaminases that act on RNA (Nishikura 2010; Orlandi et al. 2012; Wulff & Nishikura 2010), and C-to-U differences arise from deamination of cytidine to uridine by ABOBEC1, a member of the AID/APOBEC gene family (Conticello 2012; Keegan et al. 2001; Smith et al. 2012; Wedekind et al. 2003). Inosine is recognized by the translational machinery as guanosine, and thus, A-to-I RNA editing by ADAR is functionally equivalent to A-to-G changes. Discovered initially for its ability to unwind RNA duplexes through adenosine deamination (Bass & Weintraub 1988), ADAR is now recognized for its involvement in post-transcriptional A-to-I editing of various double-stranded RNA substrates, playing a role in proteome diversification (Pullirsch & Jantsch 2010), regulation of gene expression through alternative splicing (Laurencikiene et al. 2006; Rueter et al. 1999; Schoft et al. 2007) and RNA interference (Alon et al. 2012; Borchert et al. 2009; Kawahara et al. 2008; Kawahara et al. 2007; Liang & Landweber 2007; Nishikura 2006; Reid et al. 2008; Yang et al. 2006). Early insights into the function of ADAR came from reports that a glutamine-to-arginine codon change in glutamate receptors controls ion flow in mouse brain (Sommer et al. 1991). Subsequent studies have demonstrated that ADAR likewise directs amino acid changes in neurotransmitters of other organisms (Hoopengardner et al. 2003; Rosenthal & Bezanilla 2002; Seeburg & Hartner 2003) and thus plays an important role in nervous functions (Jepson & Reenan 2008; Jepson et al. 2012) and developmental

and psychiatric disorders (Bhogal et al. 2011; Grohmann et al. 2010; Sawada et al. 2009; Silberberg et al. 2012; Tan et al. 2009; Zhu et al. 2012). Furthermore, RNA editing by ADAR is necessary for proper development and normal life as ADAR-deficient invertebrates exhibit behavioral defects (Palladino et al. 2000; Tonkin et al. 2002), ADAR1-knockout mice are embryonic lethal (Q. Wang et al. 2000), and ADAR2-knockout mice are viable but die prematurely (Higuchi et al. 2000).

Within the past decade, different approaches have been developed to identify A-to-I editing sites in a global manner. Prior to the development of these genome-wide screens, only a few genes were known to be targets of ADAR (Bass 2002; Hoopengardner et al. 2003; Morse & Bass 1999). An initial study aimed at global discovery of RNA editing sites searched in large databases of expressed sequence tags and found more than 12,000 sites in 1,600 genes, increasing the number of previously known editing sites in humans by more than two orders of magnitude (Levanon et al. 2004). This work revealed that the majority of A-to-I editing sites in humans are located within noncoding regions of the gene, typically in *Alu* elements, a class of short interspersed elements (SINEs) unique to primates that comprises approximately 10% of the human genome (Batzer & Deininger 2002). With the emergence of next-generation sequencing technology, recent studies have used RNA-Seq to identify editing sites in a global manner across different cell types and organisms (Bahn et al. 2012; Chepelev 2012; Ju et al. 2011; J. B. Li et al. 2009; Park et al. 2012; Peng et al. 2012; Ramaswami et al. 2012). These studies confirm the notion that A-to-I RNA editing in mammalian systems is widespread, with estimates ranging from approximately 1,000 to over 400,000 in humans (Ramaswami et al. 2012). These studies also show that in humans, the vast

majority of editing sites are within noncoding regions (mainly the 3'UTR and introns) and that codon changes in protein sequences are rare (Kleinman et al. 2012). Expanding the search for editing events beyond processed mRNA transcripts, some next-generation sequencing studies have investigated RNA editing in different species of RNA, such as small RNA, lending insight into crosstalk between the ADAR and RNA intereference pathways (Alon et al. 2012; Warf et al. 2012; D. Wu et al. 2011). Other RNA-Seq studies have focused on the temporal aspect of editing by sequencing nascent RNA, discovering that the majority of A-to-I editing in *Drosophila* occurs cotranscriptionally (Rodriguez et al. 2012).

The second type of RNA editing in metazoa is C-U deamination by APOBEC1 (Keegan et al. 2001). Discovered for its role in producing two distinct forms of apolipoprotein B: apoB100 and the shorter isoform apoB48, which results from the conversion of a glutamine codon to a stop codon through C-to-U deamination in *apo*B mRNA (S. H. Chen et al. 1987; Hospattankar et al. 1987; Powell et al. 1987; Smith et al. 1997), APOBEC1 is a zinc-dependent cytidine deaminase that achieves editing site specificity through local sequence motifs (Backus & Smith 1994; Hersberger & Innerarity 1998; Hersberger et al. 1999; Shah et al. 1991) and the RNA-binding "APOBEC1 complementing factor" ACF (Mehta et al. 2000). *Apo*B is a lipoprotein that plays a critical role in the transport of cholesterol and triglycerides in the plasma (Chan 1992). In humans, C-to-U editing of *apo*B mRNA occurs only in the small intestine but not the liver (Greeve et al. 1993). ApoB100 expression in the liver plays a critical part in the removal of low-density lipoproteins (LDL) through the interaction of the carboxyl terminus of apoB100 with LDL receptors (Innerarity et al. 1990). This function of

apoB100 in modulating LDL levels is important in the development of atherosclerosis, as high LDL levels are a major risk factor for coronary heart disease (Ross 1995). In contrast, apoB48, which is produced in the small intestine and lacks the carboxyl terminus of apoB100, aids in the synthesis and secretion of chylomicrons (Kane et al. 1980). In contrast to widespread A-to-I RNA editing, only a few additional targets of APOBEC1 have been identified apart from *apo*B mRNA, such as the editing of glycine receptors (Legendre et al. 2009; Meier et al. 2005), *NAT1* (novel APOBEC target) mRNA (Yamanaka et al. 1997), and the neurofibromatosis *NF1* mRNA (Skuse et al. 1996). Recent genome-wide studies using RNA-Seq have uncovered 32 more mRNA target sites of APOBEC1, all of which are in AU-rich segments of the 3' untranslated regions (3' UTR) of gene transcripts (Rosenberg et al. 2011). The localization of these editing sites to the 3' UTR of transcripts may alter binding sites for RNA-binding proteins, abolish or create miRNA seed sequences, or affect additional post-transcriptional processes such as polyadenylation, subcellular localization, or translational efficiency, although further research into the biological impact of these editing events remains to be done.

1.4.3 Other types of RNA-DNA sequence differences

In addition to A-to-I editing by ADAR and C-to-U editing by APOBEC1 in metazoa and other types of RNA editing processes in other organisms, many researchers have reported the discovery of RNA-DNA sequence differences (RDDs) with unknown mechanisms. We refer to these RDDs as noncanonical RDD events. Initially, knowledge of these RDDs arose from observations of discrepancies between cloned transcripts of DNA versus cDNA. Examples include a C-to-U difference in the *WT1* gene implicated in Wilms' tumor pathology (Sharma et al. 1994), a U-to-A difference in the gene coding for

α-galactosidase (Novo et al. 1995), and various types of RDDs in the β amyloid precursor protein and ubiquitin-B protein genes of Alzheimer's and Down patients (van Leeuwen et al. 1998), the transcobalamin II gene (Qian et al. 2002), the androgen receptor gene (Martinez et al. 2008), and in HIV mRNA of chronically infected H9 cells (Bourara et al. 2000). In addition to these isolate examples in individual genes, genome-scale analyses of expressed sequence tags revealed elevated levels of non-random RDDs in cancer samples versus those of normal tissue (Bianchetti et al. 2012; Brulliard et al. 2007).

With recent advances in sequencing and informatics, several groups have utilized next-generation sequencing technology to perform systematic and genome-wide comparisons of DNA and RNA sequence. In 2011, Cheung and colleagues obtained RNA-Seq data on a group of 27 unrelated individuals in the Centre d'Etude du Polymorphisme Humain (CEPH) collection (Dausset et al. 1990) for whom low-coverage whole-genome sequence information is available and found widespread occurrences of all 12 types of RDDs, the majority of which cannot be explained by known mechanisms (M. Li et al. 2011). A few of these noncanonical RDD events were further validated by Sanger sequencing and also shown to be translated into protein by mass spectrometry. Shortly thereafter, other groups reported similar findings of noncanonical RDD events using cell lines derived from different ethnic groups (Ju et al. 2011) and tissue types (Bahn et al. 2012).

However, several research groups subsequently challenged the existence of noncanonical RDD events reported by Cheung and colleagues, contending that the vast majority (greater than 90%) of the observed sequence discrepancies are caused by technical artifacts due to misalignment, sequencing, and genotyping error (Kleinman &

Majewski 2012; W. Lin et al. 2012; Peng et al. 2012; Pickrell et al. 2012; Schrider et al. 2011). To correct for these various sources of error, several research groups developed different filtering criteria and computational techniques for the accurate identification of RDDs (Kleinman et al. 2012; Peng et al. 2012; Ramaswami et al. 2012). These pipelines aim to account for misalignment in challenging regions of the genome such as repetitive sequences, pseudogenes, homologous sequences, and exon-exon junctions, and address experimental noise such as strand and positional bias in sequencing error. The results from these stringent computational pipelines indicate that A-to-G events comprise approximately 80 to 90% of all RDDs identified and are able to be reproducibly validated using Sanger sequencing technology at rates of about 90% across various studies (Peng et al. 2012; Ramaswami et al. 2012). In contrast, noncanonical RDD events vary from not being able to be confirmed by Sanger sequencing (Ramaswami et al. 2012) to validation rates of approximately 50% (Peng et al. 2012). As such, the debate over the existence and prevalence of noncanonical RDDs in humans continues persists, with evidence suggesting that they do occur albeit at a low frequency. Further mechanistic studies are needed to explain their origin and functional impact on the cell.

1.5 Summary

The study of the transcriptome, or the complete set of transcribed RNA species in the cell at any given time under a particular set of conditions, is critical for understanding how expression of information in the genome is translated into overall phenotypic variation. Recent advances in sequencing technology have enabled genome-wide studies of RNA output at unprecedented depths and resolution. In Chapter 2, we explore the use of next-generation sequencing technology in global analyses of gene expression. In particular, we obtain deep RNA sequence information of cultured human B-cells to quantify the expression levels of all genes and their transcripts and assess the depths of sequencing necessary for sequencing-based studies. We provide the results of our transcriptome analysis as a public resource for others interested in the expression and structure of genes in human B-cells. Next, in Chapter 3, we examine systematic sequence differences between RNA and DNA of the same individual, or sequence alterations that may result from processes such as RNA editing, transcriptional infidelity, or other unknown mechanisms. In light of the ongoing debate over the prevalence and existence of RNA-DNA sequence differences (RDDs) mediated by unknown mechanisms, we perform *in silico* and experimental analyses to evaluate the precision and power of next-generation sequencing technology and associated computational methods in the detection of RDDs in chapter 4. Using various bioinformatics algorithms and filtering methods to control for false positives, we develop a computational pipeline for the accurate identification of RDDs. Lastly, in chapter 5, we briefly evaluate the genetic basis for RDDs using RNA-Seq data. In particular, we first evaluate the extent to which the levels of RDDs, or the percentage of transcripts at a given site that differ from the underlying

genomic template sequence, vary across individuals and furthermore, the degree to which this variation is genetically determined.

In summary, this thesis provides a resource for those interested in the study of transcriptomes using RNA-Seq technology and computational methods. In particular, this work lends insight into the analysis of gene expression and systematic detection of sequence variants between RNA and DNA genome-wide. The results from the gene expression profiling of human B-cells are available on the UCSC Genome Browser, and the computational pipeline for the identification of RDDs is outlined in the thesis. Lastly, the work on the heritability of RDDs lays the foundation for future genetic analyses of the determinants influencing systematic sequence differences between DNA and RNA sequence.

Chapter 2. RNA-Sequence Analysis of Human B-Cells

2.1 Abstract

RNA-Sequencing (RNA-Seq) allows for quantitative measurement of expression levels of genes and their transcripts. In this chapter, we analyzed the transcriptome of cultured human B-cells. In particular, we sequenced complementary DNA fragments (cDNA) derived from human lymphoblastoid cell lines and obtained 879 million 50 base-pair reads comprising 44 Gigabases of sequence. The results enabled us to evaluate the expression profile of B-cells and to assess experimental parameters for sequencing-based studies. Overall, we identified 20,766 genes and 67,453 of their alternatively-spliced isoforms. More than 90% of genes with multiple exons are alternatively-spliced, and for most genes, one isoform is expressed predominantly. We observed that while chromosomes differ greatly in gene density, the percentage of transcribed elements in each chromosome is less variable. In addition, genes involved in related biological processes are expressed at more similar levels than genes with different functions. Besides the analysis of gene expression levels, we also used the data to investigate the effect of sequencing depth on gene expression measurements. While 100 million reads are sufficient to detect most expressed genes and transcripts, about 500 million reads are needed to measure accurately their expression levels. We provide examples in which deep sequencing is needed to determine the relative abundance of genes and their isoforms. With data from 20 individuals and about 40 million sequence reads per sample, we uncovered only 21 alternatively-spliced, multi-exon genes that are not in existing databases; this result suggests that at this sequence coverage, we can detect most known genes. Results from this project are available on the UCSC Genome Browser to allow

readers to study the expression and structure of genes in human B-cells. The majority of this work is adapted from previously published results (Toung et al. 2011).

2.2 Introduction

Cellular phenotypes are determined in large part by gene expression. Thus, a comprehensive catalog of gene transcripts, their structures, and abundance is critical for understanding how gene expression influences phenotypic manifestations. In the past, microarrays (DeRisi et al. 1996; Fodor et al. 1993) have been the predominant method for gene expression studies because of their ability to probe thousands of transcripts simultaneously. Although these hybridization-based approaches are high throughput, they are subject to biases and limitations such as the reliance on existing gene models and potential for cross-hybridization to probes with similar sequences. To overcome some of these restrictions, genomic tiling arrays and other approaches such as serial analysis of gene expression (Velculescu et al. 1995) and massively parallel signature sequencing (Brenner et al. 2000) have been developed.

RNA-sequencing (RNA-seq) is a relatively new approach for analyzing gene expression; it provides digital readouts for mapping and quantifying transcriptomes (Bentley et al. 2008; Lister et al. 2008; Mortazavi et al. 2008; Nagalakshmi et al. 2008; Wilhelm et al. 2008). The method involves isolating a population of RNA, converting it to a library of cDNA fragments with adaptors attached, and sequencing the cDNA library to obtain short sequences typically 30 to 400 nucleotides in length. The short reads are then mapped to a reference genome or assembled de novo. The expression level for a gene can subsequently be determined by counting the number of reads that aligned to its exons. RNA-Seq studies of model organisms (Cloonan et al. 2008; Mortazavi et al. 2008) have revealed unknown aspects of transcriptomes through refinement of transcriptional start sites, discovery of 3’ UTR heterogeneity, and identification of novel upstream open

reading frames. Global surveys of alternative splicing show that nearly 95% of all multi-exon genes in humans undergo alternative splicing events (Pan et al. 2008). Motivated by the ability of RNA-Seq technology to study gene expression, we sequenced the transcriptomes of human B-cells that are part of the HapMap (International HapMap Consortium 2005) and 1000 Genomes Projects (1000 Genomes Project Consortium 2010). We generated 44 Gigabases of sequence to address several questions. First, we analyzed the gene expression landscape of human B-cells by identifying expressed transcripts and quantifying their expression levels. Second, we examined how sequencing depth affects the detection and quantification of genes and their isoforms. Lastly, we evaluated the potential of RNA-Seq to uncover transcribed fragments that are not in existing gene annotation databases.

2.3 Results

2.3.1 Dataset

We sequenced the mRNA population of cultured human B-cells from 20 unrelated individuals from the Center d'Étude du Polymorphisme Humain (CEPH) collection (Dausset et al. 1990). For each sample, we obtained 44 ± 8 million 50-bp reads (mean ± standard deviation) (see Materials and Methods). For most of the analysis, we pooled the sequences to create a dataset comprising 879 million reads or 44 Gigabases of sequence; we refer to this dataset as the "879-million-read" dataset.

We mapped the sequence reads to the reference human genome sequence (NCBI 36.1/hg18 assembly) using TopHat (Trapnell et al. 2009) and Bowtie (Langmead et al. 2009). Then, we assembled the alignments into gene transcripts and calculated their relative abundances using Cufflinks (Trapnell et al. 2010). We conducted two analyses: First, we provided Cufflinks with Gencode (v. 3c) (Harrow et al. 2006) gene annotations, and second, we did not use any gene annotations to find unknown gene models. We restricted our first analysis to levels 2 and 3 Gencode genes that are annotated as "protein coding" or "processed transcript"; in this study, we refer to this set of gene models as "Gencode". To investigate the effect of sequencing depth on various expression profiling measurements, we created smaller subsets of our pooled data set, analyzing depths of 1 to 9 million reads (in intervals of 1 million reads), 10 to 90 million reads (in intervals of 10 million reads), and 100 to 700 million reads (in intervals of 100 million reads).

2.3.2 Alignment results

In the 879-million-read dataset, 80% of the reads aligned to the human genome, of which 84% aligned to unique locations in the genome (Table 2.1). Fourteen percent of the mapped reads aligned to two to five locations in the genome, and less than 2% aligned to six or more locations. We excluded all reads mapping to six or more locations from our analyses. Although less than 3% of the human genome is composed of exons, 83% of our uniquely mapped reads overlap Gencode exons. These results confirm that our poly(A)+ RNA samples are highly enriched for exonic sequences. We also studied fractions of the 879-million-read data set and found that the percentage of total reads aligning to the human genome increases proportionally with sequencing depth for input sizes smaller than 200 million reads, after which the value remains constant. With 1 million reads, 75% of the reads aligned to the genome; in contrast, 80% of the reads aligned with 200 million reads (Table 2.1). Lastly, we found that 84% of the aligned reads mapped to unique locations across all sequencing depths.

Table 2.1 RNA-Seq alignment results. Alignment statistics for the pooled 879-million-read dataset and smaller subsets are depicted.

Number of Reads Sequenced (Millions)	Number of Reads Aligned	Number of Reads Aligned Uniquely	Number of Reads Aligned to 2-5 Locations	Number of Reads Aligned to 6+ Locations
1	753,387 (75%)	634,176 (84%)	106,142 (14%)	13,069 (2%)
2	1,515,873 (76%)	1,277,350 (84%)	212,041 (14%)	26,482 (2%)
3	2,282,729 (76%)	1,924,611 (84%)	318,303 (14%)	39,815 (2%)
4	3,050,813 (76%)	2,573,536 (84%)	424,363 (14%)	52,914 (2%)
5	3,819,893 (76%)	3,222,534 (84%)	531,167 (14%)	66,192 (2%)
6	4,590,591 (77%)	3,873,017 (84%)	638,218 (14%)	79,356 (2%)
7	5,361,450 (77%)	4,524,251 (84%)	744,318 (14%)	92,881 (2%)
8	6,133,403 (77%)	5,176,244 (84%)	850,944 (14%)	106,215 (2%)
9	6,905,747 (77%)	5,828,087 (84%)	958,025 (14%)	119,635 (2%)
10	7,678,072 (77%)	6,480,455 (84%)	1,064,761 (14%)	132,856 (2%)
20	15,652,328 (78%)	13,174,570 (84%)	2,210,319 (14%)	267,439 (2%)
30	23,631,519 (79%)	19,873,435 (84%)	3,356,152 (14%)	401,932 (2%)
40	31,613,156 (79%)	26,575,527 (84%)	4,500,716 (14%)	536,913 (2%)
50	39,596,782 (79%)	33,275,797 (84%)	5,648,880 (14%)	672,105 (2%)
60	47,576,773 (79%)	39,975,008 (84%)	6,795,093 (14%)	806,672 (2%)
70	55,555,326 (79%)	46,672,758 (84%)	7,940,864 (14%)	941,704 (2%)
80	63,533,073 (79%)	53,366,429 (84%)	9,090,087 (14%)	1,076,557 (2%)
90	71,508,722 (79%)	60,058,715 (84%)	10,238,900 (14%)	1,211,107 (2%)
100	79,484,383 (79%)	66,749,857 (84%)	11,388,069 (14%)	1,346,457 (2%)
200	159,629,770 (80%)	133,939,976 (84%)	22,998,295 (14%)	2,691,499 (2%)
300	239,550,734 (80%)	200,876,315 (84%)	34,636,304 (14%)	4,038,115 (2%)
400	319,545,150 (80%)	267,789,138 (84%)	46,370,553 (15%)	5,385,459 (2%)
500	399,305,262 (80%)	334,443,831 (84%)	58,125,933 (15%)	6,735,498 (2%)
600	477,748,475 (80%)	399,783,258 (84%)	69,884,210 (15%)	8,081,007 (2%)
700	557,037,521 (80%)	465,915,594 (84%)	81,691,919 (15%)	9,430,008 (2%)
879	700,391,492 (80%)	585,460,567 (84%)	103,112,278 (15%)	11,818,647 (2%)

2.3.3 Expression landscape of human B-cells

Using all of our sequence reads, we estimated the expression levels of genes in our B-cells. Expression levels are measured in "fragments per kilobase of exon model per million mapped reads" (FPKM) (Trapnell et al. 2010), and the expression level for a gene is the sum of the FPKM values of its isoforms. The distribution of gene expression values is right-skewed (Figure 2.1); the median and mean FPKM values are 26 and 338, respectively. Although we do not wish to use an arbitrary FPKM threshold to determine whether a transcript is expressed, analysis of all transcripts with expression levels greater than zero will include FPKM values that are very close to zero (bottom fifth percentile of transcript FPKM values = 0.003). Thus, we set an FPKM value of 0.05 as the lower bound in our subsequent analyses. Using this criterion, we detected 20,776 genes and 67,453 alternatively-spliced transcripts in our B-cells. For the majority (75%) of these transcripts, there are sequence reads that cover at least one-quarter of their exons. The expression of these transcripts is supported by RNA polymerase II binding and active chromatin marks such as H3K27ac or H3K4me3 (Figure 2.2) (Rosenbloom et al. 2010).

Figure 2.1 Distribution of gene expression levels in human B-cells. A histogram depicting expression levels for genes as defined by Gencode in units of FPKM. The distribution is skewed right; the median and mean FPKM values are 26 and 338, respectively. The main figure shows genes with FPKM values less than 1000, and the inset shows genes with FPKM values greater than 1000.

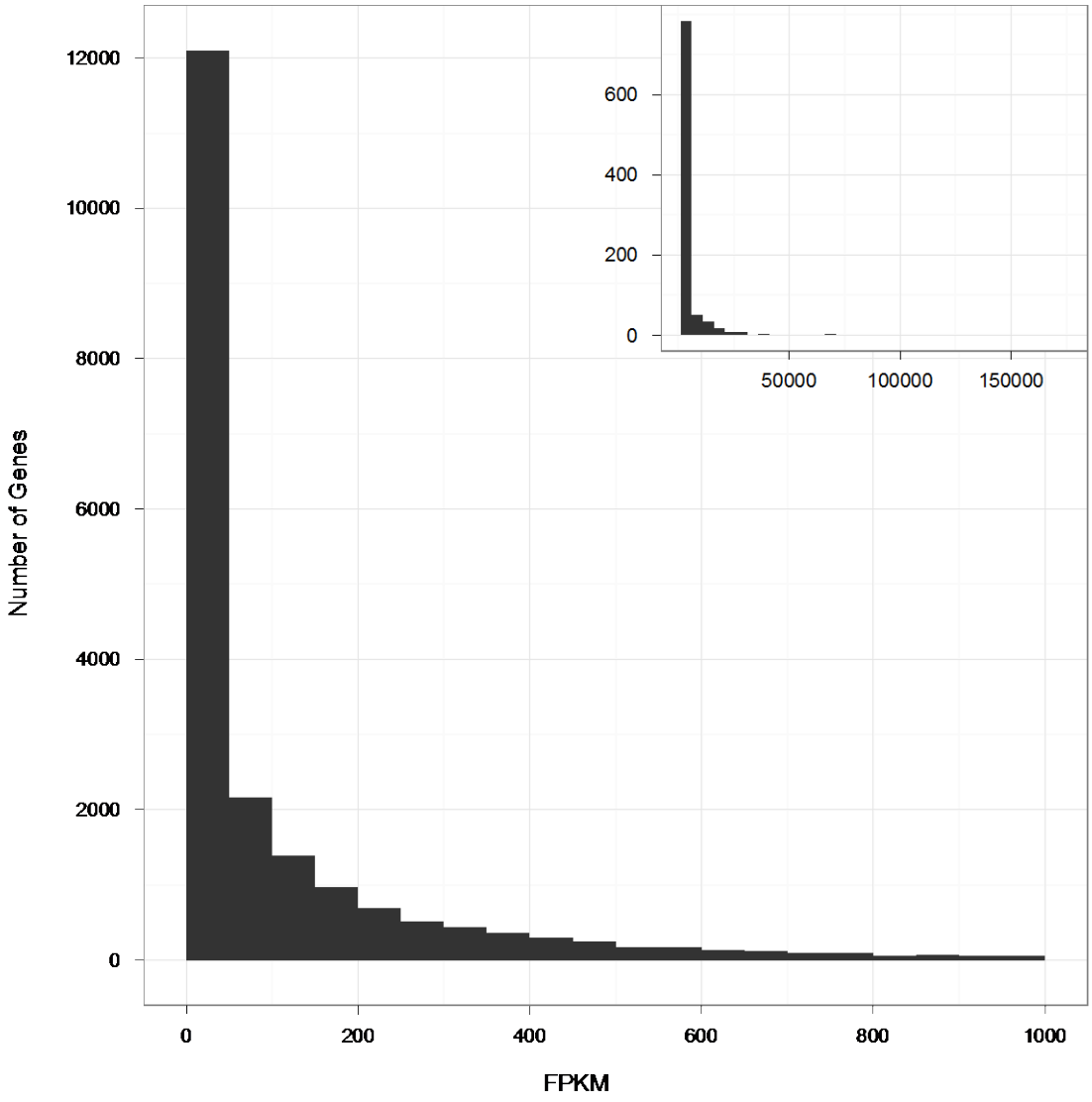
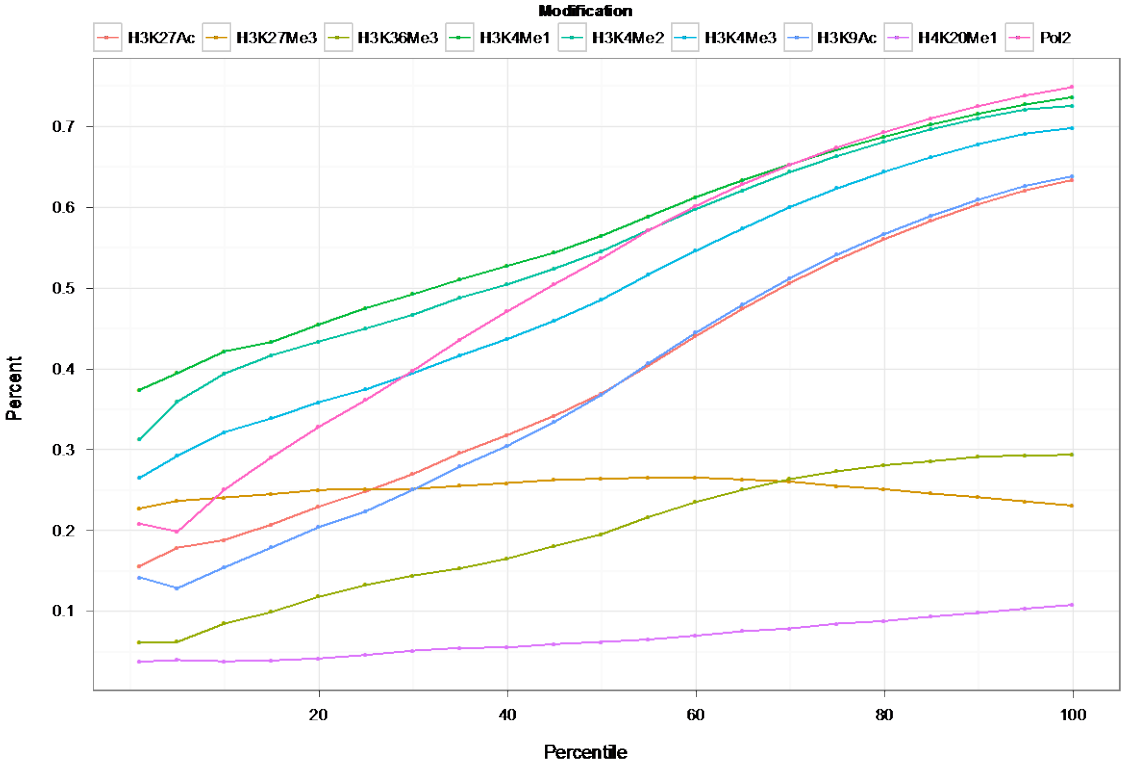


Figure 2.2 Percent of genes overlapping various ENCODE gene regulation tracks. The percentage of genes overlapping RNA Polymerase II (Pol2) binding sites and various chromatin marks (y-axis) for various FPKM percentiles (x-axis). For example, the percentage of genes with FPKM values greater than the 30th FPKM percentile that overlap Pol2 binding sites is 40%. We found that with increasing FPKM percentiles, the percentage of genes overlapping Pol2 binding sites increases. We saw similar patterns for chromatin marks associated with enhancers, promoters, and active transcription such as H3K27Ac, H3K4Me1, H3K4Me2, H3K4Me3, H3K9Ac. In contrast, the percentage of genes overlapping the repressive chromatin mark H3K27Me3 decreased with increasing FPKM thresholds.



We surveyed the expression landscape across chromosomes by determining the fraction of genes that are expressed within 1-Mb intervals (Figure 2.3) and analyzing the relationship between gene density and percentage of genes transcribed for each chromosome (Figure 2.4). The average gene density is 10 genes per Mb (standard deviation = 4.8), and the average percentage of genes transcribed for each chromosome is 71% (standard deviation = 12%). We found that while chromosomes varied greatly with respect to gene density, they varied much less in the proportion of genes that are expressed. For example, while chromosome 19 is six times denser in gene content than chromosome 18, 87% and 82% of genes on chromosome 19 and chromosome 18 are expressed.

Figure 2.3 Distribution of expressed genes by chromosome. For each chromosome, the number (*y*-axis) of Gencode genes residing in 1-Mb intervals along the chromosome (*x*-axis depicts physical distance in megabases) is depicted. The number of genes that are expressed (FPKM ≥ 0.05) is colored in red, and the number of genes that are not expressed is shaded blue.

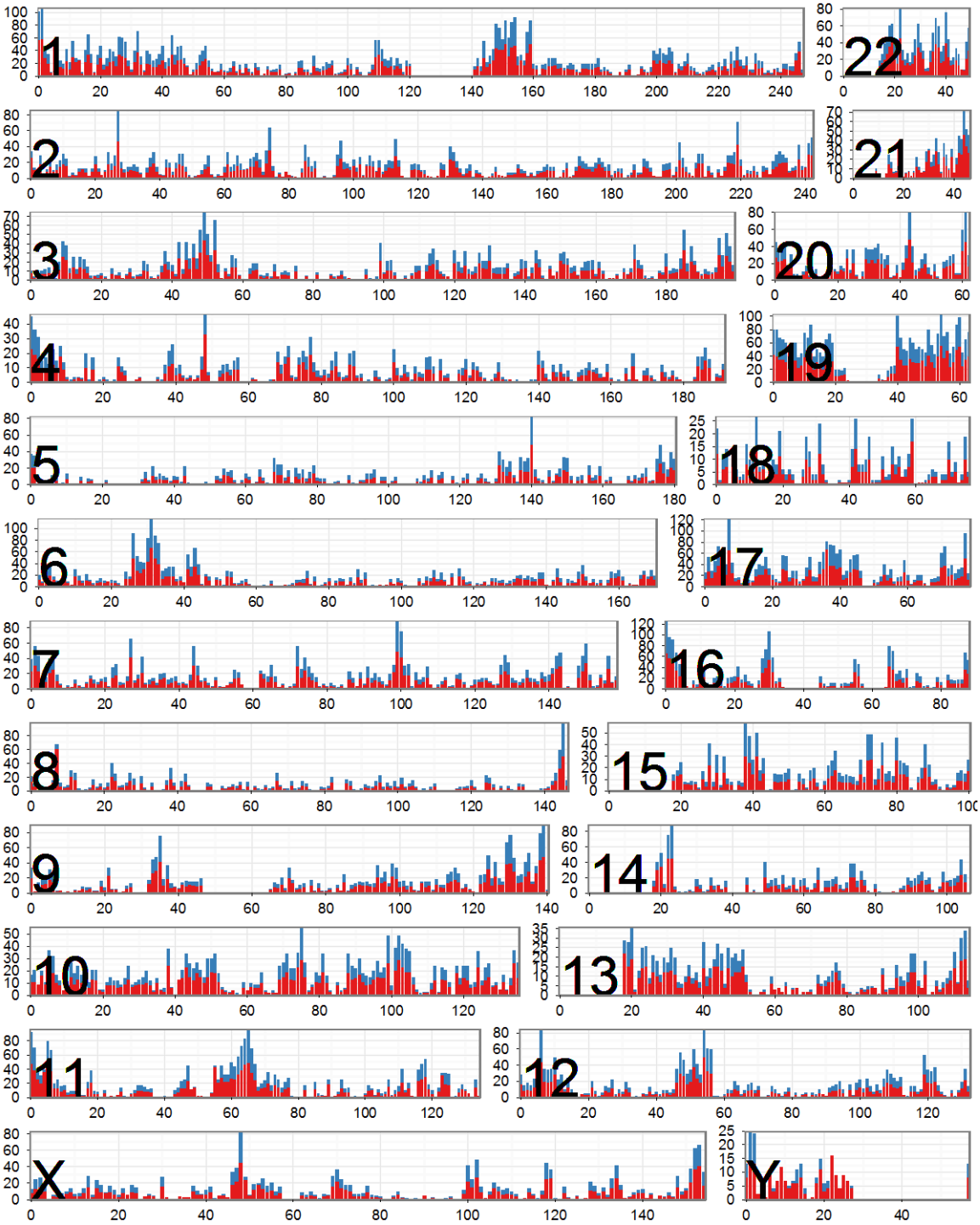
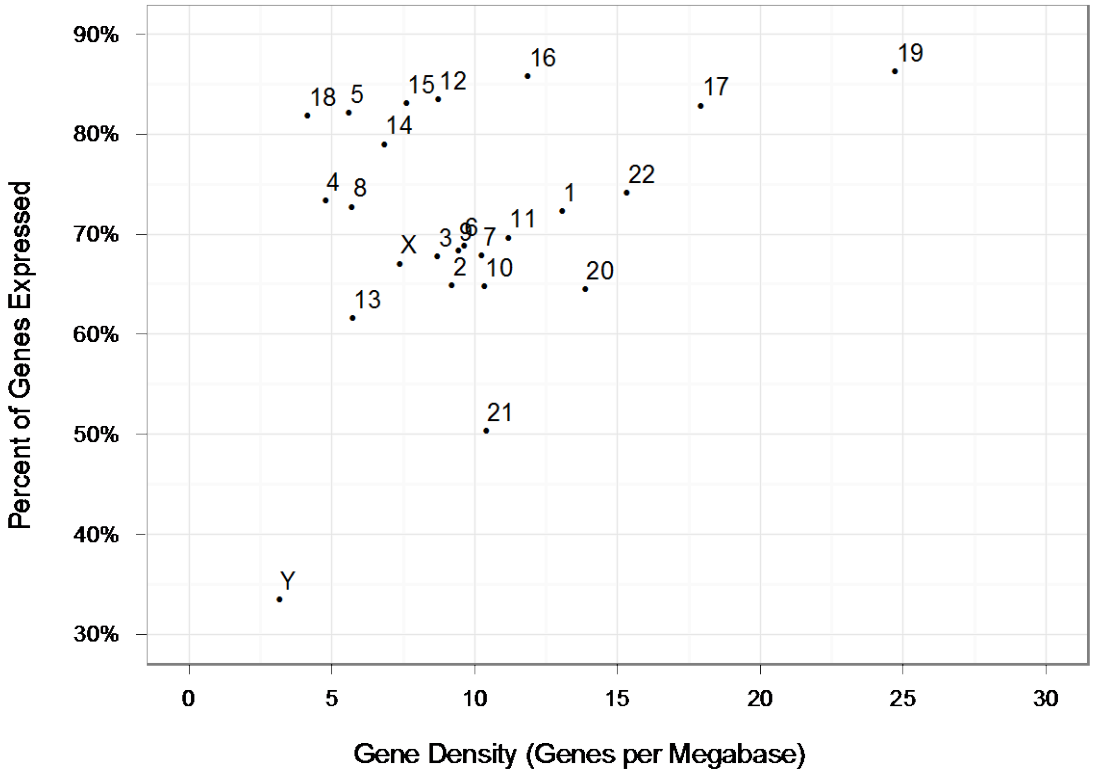


Figure 2.4 Gene density versus percentage of genes transcribed. For each chromosome, the percentage of genes expressed (*y*-axis) is shown versus the gene density (*x*-axis) ($R = 0.34$). We see that while genes vary greatly in their gene density, they differ less in the percentage of genes expressed.

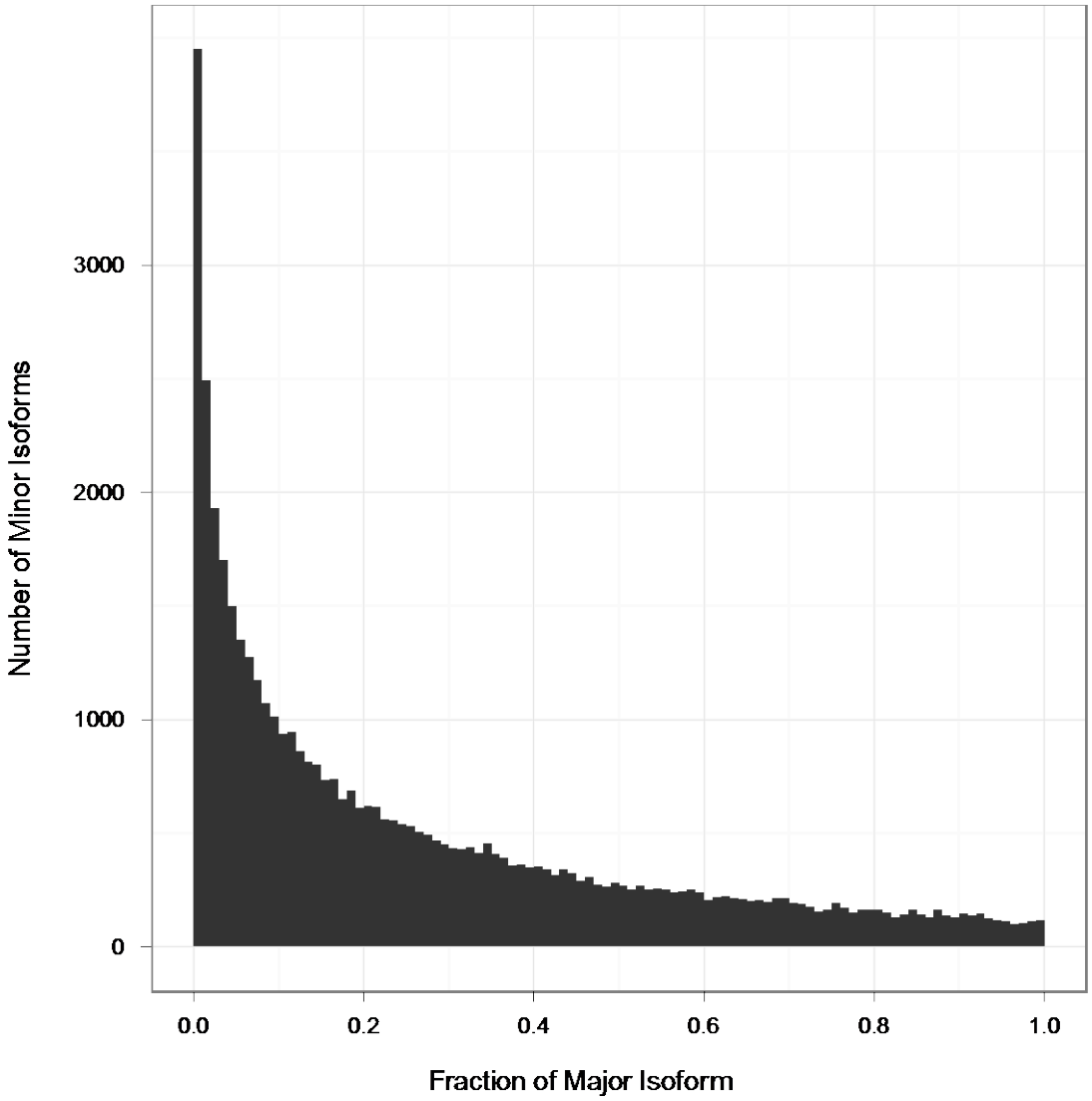


We classified genes into groups based on their FPKM values: low expression (bottom 25th percentile; FPKM $\leq$ 2.3), medium expression (middle 50th percentile; 2.3 < FPKM $\leq$ 163), and high expression (top 25th percentile; FPKM > 163). Gene Ontology (GO) analysis (Ashburner et al. 2000) revealed that low-expressing genes are enriched for processes relating to cell adhesion ($P = 2.9 \times 10^{-20}$) and ion transport ($P = 1.1 \times 10^{-15}$). For medium-expressing genes, genes involved in transcription ($P = 2.4 \times 10^{-31}$) were found to be overrepresented. Lastly, we found high-expressing genes to be enriched in processes such as translation ($P = 3.1 \times 10^{-70}$), RNA processing (2.2×10^{-70}), and RNA splicing (5.3×10^{-56}). We did not find functional categories that were enriched in all three groups, suggesting that genes within a particular process are expressed at similar levels.

2.3.4 Alternative splicing activity in human B-cells

We assessed the degree of alternative splicing activity in B-cells and found that 94% of multi-exon genes express two or more spliced forms. This number is consistent with the estimate by Burge and colleagues (E. T. Wang et al. 2008) that greater than 90% of human genes across diverse tissue types express multiple isoforms. For genes with two or more expressed isoforms, we analyzed the relative abundance of each of the alternatively-spliced transcripts. We considered the transcript with the highest FPKM value as the "major" isoform and all other transcripts as "minor" isoforms. For every minor isoform of a gene, we calculated the ratio of its FPKM value to that of the major isoform. We found the distribution of these ratios to be right-skewed with a mean of 0.26 (median = 0.17, standard deviation = 0.26) (Figure 2.5). These results indicate that while the majority of genes have several alternatively-spliced transcripts, these isoforms are not expressed at equivalent levels. For most genes, one isoform is expressed more highly than others.

Figure 2.5 Distribution of 'fraction of major isoform' values. For genes with two or more alternatively-spliced transcripts, we considered the transcript with the highest FPKM value as the "major" isoform and all others as "minor" isoforms. For each minor isoform, we calculated its "fraction of major isoform" value to be the ratio of its FPKM value to that of the major isoform. Here we plot the distribution of "fraction of major isoform" values. The distribution is right skewed with a mean of 0.26, median of 0.17 and standard deviation of 0.26. These results suggest that most genes express minor isoforms at relatively low values compared to the major isoform.



2.3.5 Concordance of gene expression levels by RNA-Seq and microarrays

We compared our RNA-Seq data to microarray measurements performed on the same 20 unrelated CEPH individuals. The gene expression levels measured by the two methods are similar ($R = 0.59$) (Figure 2.6) and comparable to results by others. Mortazavi and colleagues measured a correlation of 0.69 in mice (Mortazavi et al. 2008), and Montgomery and colleages found a correlation of 0.80 in human cell lines (Montgomery et al. 2010). To investigate whether the digital counts of transcript abundance produced by RNA-Seq experiments offer greater dynamic range than the analog-style signals obtained from microarrays, we analyzed the expression levels for 2,597 genes for which data were available for each of the 20 individuals. For each gene, we calculated the dynamic range and the coefficient of variation. We found the dynamic range to be greater in RNA-Seq than microarray measurements: 1.78 ± 0.67 versus 1.25 ± 0.47 (mean ± standard deviation). Across the 20 individuals, the coefficient of variation values was also greater from RNA-Seq data: 0.13 ± 0.09 versus 0.052 ± 0.03(mean ± standard deviation). For the majority (90%) of the genes, the coefficient of variation is larger in the RNA-Seq data set (see examples in Figure 2.7).

Figure 2.6 Expression values from RNA-Seq and microarrays. Comparison of FPKM values (log_2-transformed) and microarray signals for the 2597 genes detected by both platforms in 20 unrelated CEPH individuals. For each gene, we plotted the average expression values across the 20 individuals. Individual points are colored grey with transparency; darker dots represent overlapping genes.

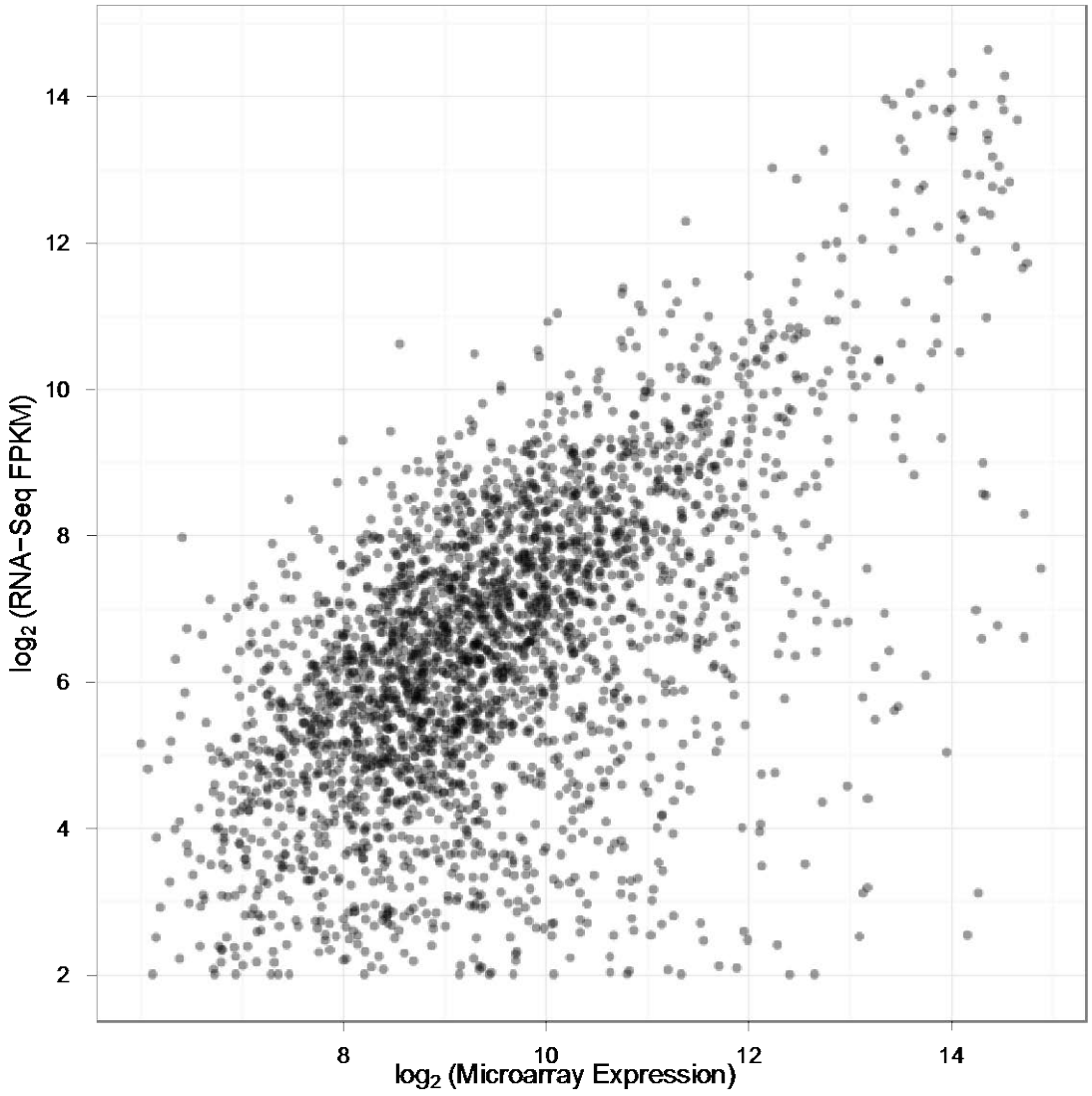
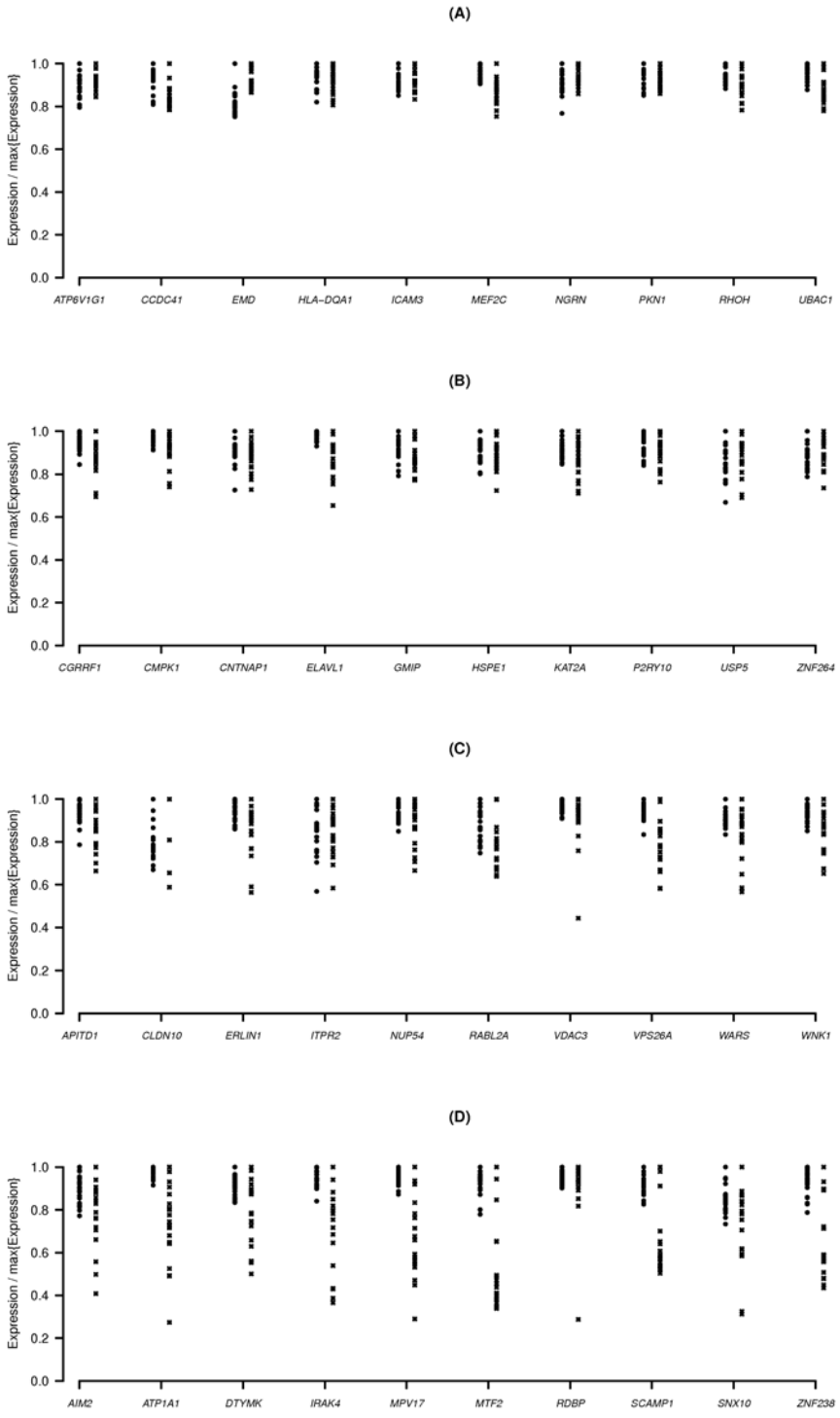


Figure 2.7 RNA-Seq and microarray expression values versus coefficient of variation. Expression levels (scaled to facilitate comparison) as measured by microarrays (left) and RNA-Seq (right) for various randomly chosen genes. Genes are grouped by whether the coefficient of variation lies in the (A) first quartile, (B) second quartile, (C) third quartile and (D) fourth quartile.



2.3.6 Effect of sequencing depth on RNA-Seq measurements

In designing an RNA-Seq study, a parameter of interest is the sequencing depth needed to address various questions. To assess the relationship between sequencing depth and expression levels, we divided our 879 million 50-bp read data set into smaller sets and analyzed how the detection of a gene and the measurement of its expression level varies with increasing sequencing depth.

We first assumed that our 879-million-read data set gives a comprehensive catalog of transcribed genes and then assessed how many genes are detected in fractions of those reads. We found that with 100 million reads, 81% of genes (FPKM $\geq$ 0.05) and 90% of their transcripts were detected (Figure 2.8). For each additional 100 million reads, we were able to detect on average 3% more genes and 1% more transcripts. As expected, the expression level of a gene affects how readily it is detected; for example, with 100 million reads, 80% of highly expressed genes (top 25th percentile; FPKM > 163) compared to 32% of the low expression genes (bottom 25^{th} percentile; FPKM $\leq$ 2.3) were detected (Figure 2.9).

Figure 2.8 Number of junctions, transcripts, and genes detected at different sequencing depths. The numbers of genes, transcripts, and junctions detected in the 879-million-read data set were assumed to be the "final" values. Then, the percentages of these "final" values detected at various sequencing depths were determined. For example, with 100 million reads, 76% of the junctions, 90% of transcripts, and 81% of genes were detected.

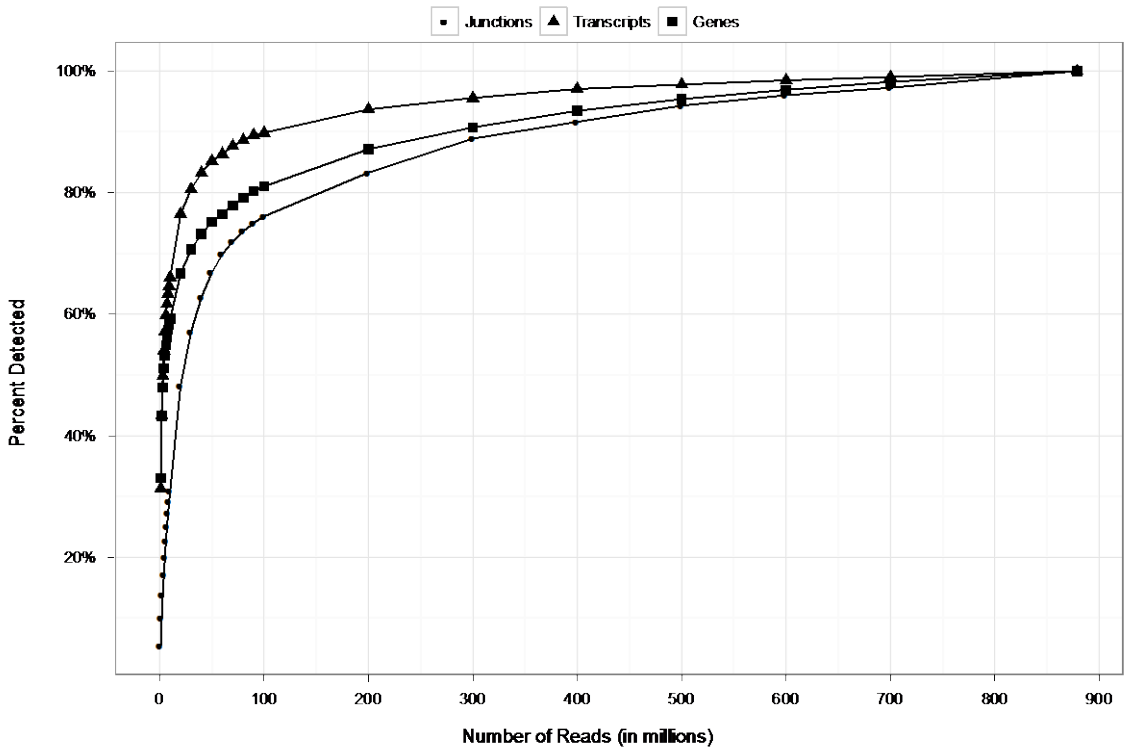
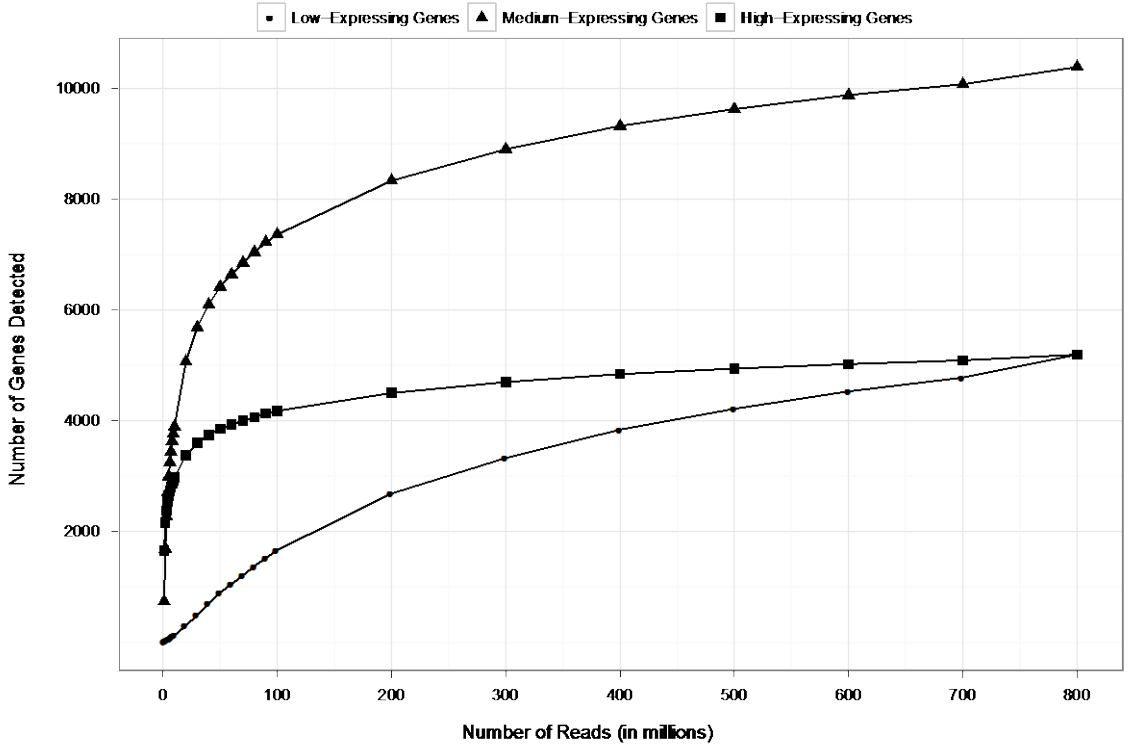


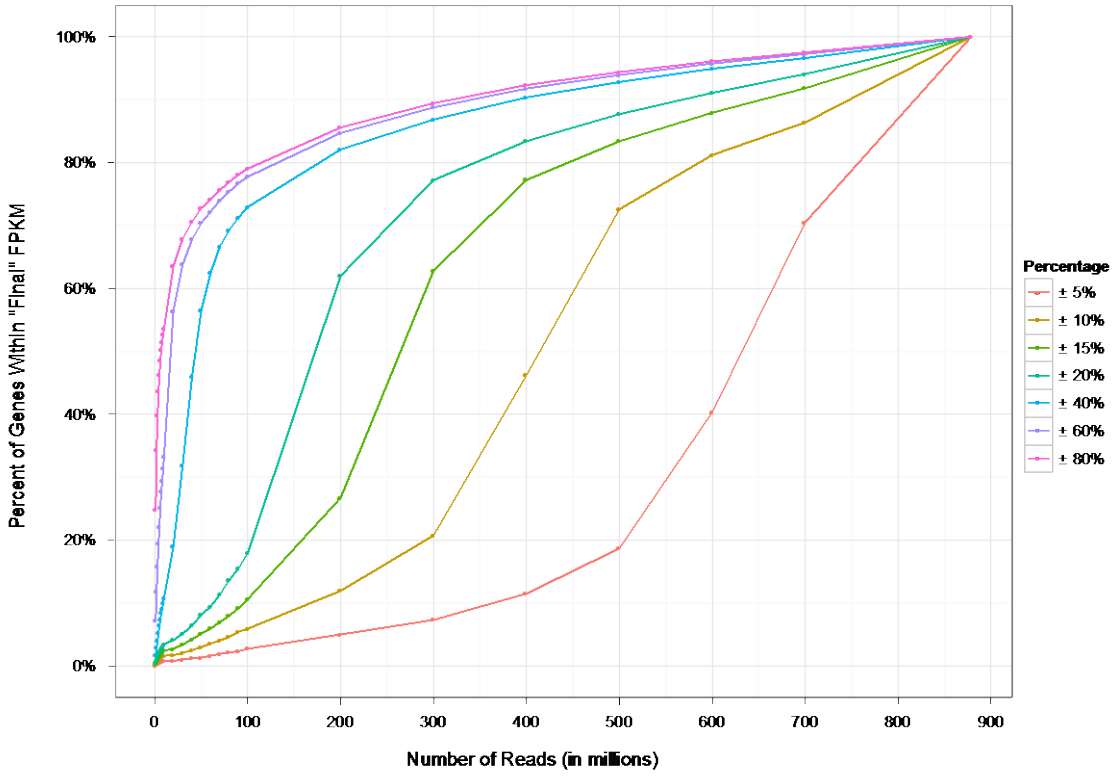
Figure 2.9 Number of genes detected at various sequencing depths. The number of genes detected at various sequencing depths is shown. Genes are grouped according to the "final" value obtained in the 879-million-read dataset: low-expressing genes (bottom 25th percentile; FPKM ≤ 2.3), medium-expressing genes (middle 50th percentile; 2.3 > FPKM ≤ 163), and high-expressing genes (top 25th percentile; FPKM > 163).



The detection of splice junctions is important as they are necessary for isoform assembly and quantification. Of the 269,155 Gencode junctions, 145,100 (54%) are detected in our 879-million read data set. This result is consistent with those reported by others: Blencowe and colleagues (Pan et al. 2008) detected between 128,395 and 130,854 of known RefSeq junctions in diverse human tissues; Pritchard and colleagues (Pickrell et al. 2010) detected 170,293 junctions supported by spliced ESTs from GenBank in B-cells. With 100 million reads, 76% of the 145,100 junctions were detected, after which on average 4% more junctions were detected for each additional 100 million reads (Figure 2.8).

For most studies, information beyond whether a gene is expressed or not is important – accurate expression levels are needed. To study the robustness of expression levels at various input sizes, we first assumed the expression values in our 879-million-read data set to be the "best estimates" and then analyzed the sequencing depth necessary to achieve these "final" levels (Figure 2.9). For the majority (72%) of genes with FPKM values greater than 0.05, 500 million reads are needed for their expression values to be within 10% of their final measurements. With 100 million reads, only 6% of genes have values that are within 10% of their "final" FPKM value. Furthermore, while 100 million reads is sufficient for detection of the majority of genes and transcripts, the expression levels of genes obtained at a depth of 100 million reads deviate on average from their final value by 41%. These results suggest that deep sequencing is necessary for accurate determination of the expression level of genes.

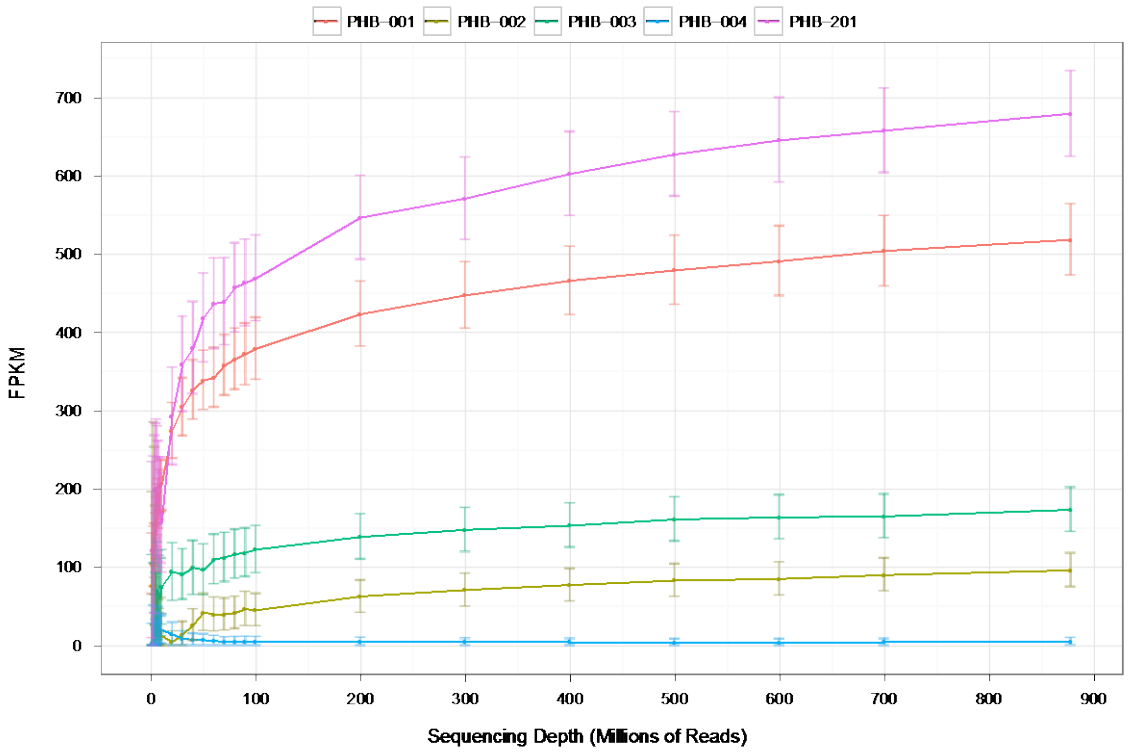
Figure 2.10 Gene expression levels at different sequencing depths. The percentages of genes that reach values within different percentages of the ‘‘final’’ level obtained at a depth of 879 million reads are depicted. With 100 million reads, only 6% of genes have FPKM measurements that are within 10% (gold line) of their “final” value compared to 72% at a depth of 500 million reads.



Next, we investigated the coverage needed to study the relative abundance of alternatively-spliced forms of genes. Again, we found that deep sequencing depths are crucial. For example, *PHB* (Figure 2.11) is a gene with five isoforms: *PHB*-001 (ENST00000300408), *PHB*-002 (ENST00000419140), *PHB*-003 (ENST00000446735), *PHB*-004 (ENST00000393345), and *PHB*-201 (ENST00000434917) with FPKM values of 519, 96, 174, 5, and 679, respectively. The expression level for the least abundant isoform (*PHB*-004) was 20% of its final FPKM at a sequencing depth of 60 million reads. However, for the other four isoforms, 200 to 400 million reads were needed to obtain expression values within 20% of their final FPKM measurements. These results are surprising as one may expect deeper sequencing to allow for better quantification of transcripts that are expressed at lower levels; however, our data suggest that it is the highly expressed isoforms whose expression levels increase with larger sequencing depths. Furthermore, with less than 200 million reads, the 95% confidence intervals reported by Cufflinks for the two most highly expressed isoforms (*PHB*-001 and *PHB*-201) overlapped each other; however, with more than 200 million reads, the confidence intervals for the five isoforms no longer intersected. Another example is *CD74*, which has three high-expressing variants: *CD74*-201 (ENST00000009530), *CD74*-202 (ENST00000353334), and *CD74*-203 (ENST00000377795) with FPKM values of 4,690, 54,745, and 2,252, respectively. While the expression level of the least-expressed isoform (*CD74*-203) was within 10% of its "final" FPKM with 20 million reads, the expression values of the other two isoforms did not reach this level until 400 million reads. Again, we see that the expression values of the highly expressed isoforms continued to increase

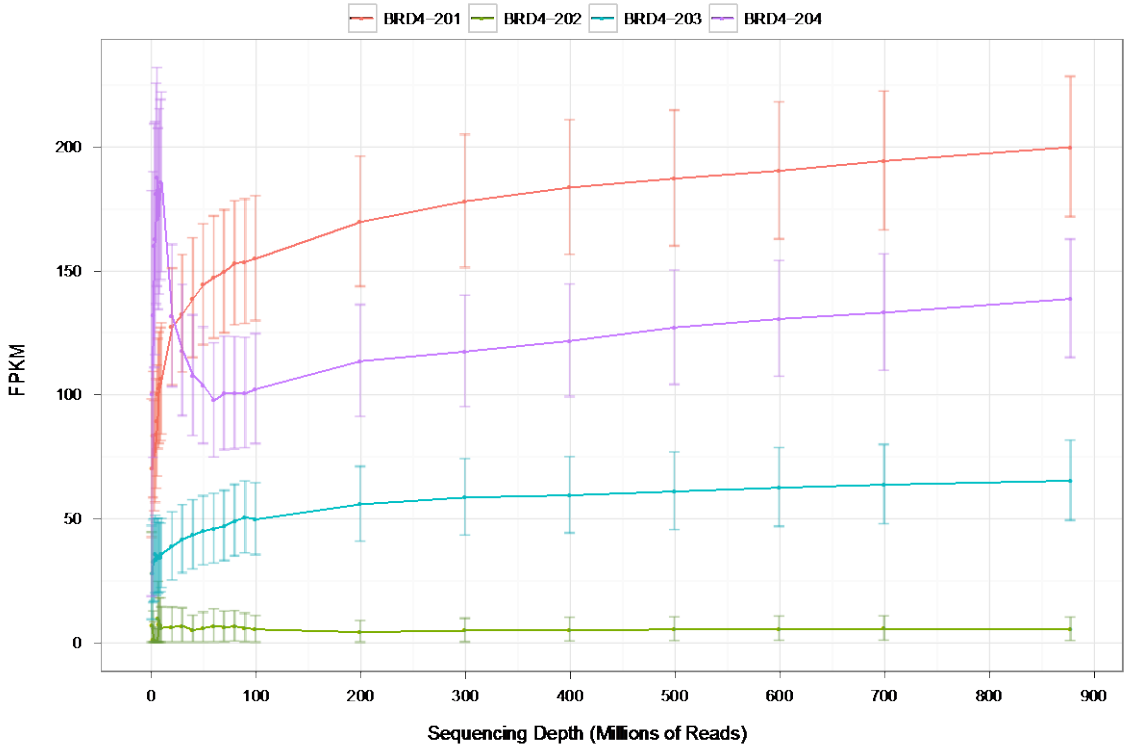
with higher sequencing depths, whereas that for the isoform with the lowest level of expression was fairly constant.

Figure 2.11 Expression levels of *PHB* versus sequencing depth. The expression levels of five alternatively-spliced transcripts of *PHB* are plotted at various sequencing depths. The least abundant isoform (blue line) of *PHB* reaches within 20% of its "final" FPKM value with only 60 million reads; however, the expression values of the other four isoforms continued to increase with deeper sequencing.



Sufficient sequence coverage is not only required for accurate estimations of expression levels; they are also necessary to determine the relative abundance of isoforms. *BRD4* is a gene with four isoforms: *BRD4*-201 (ENST00000263377), *BRD4*-202 (ENST00000360016), *BRD4*-203 (ENST00000371835), and *BRD4*-204 (ENST00000392878) with FPKM values of 200, 5, 65, and 139, respectively (Figure 2.12). At low sequencing depths, the expression level of *BRD4*-204 was overestimated, while that of *BRD4*-201was underestimated; 60 million reads were needed to show that *BRD4*-201, not *BRD4*-204, is the most highly expressed isoform. As a final example of the effect of sequencing depth on expression values, we studied relative gene expression by using two well-characterized genes − *CDKN1A*, a cyclin-dependent kinase inhibitor; and its regulator, *TP53*. The "final" FPKM values for *CDKN1A* and *TP53* were 2400 and 676, respectively; the ratio of the expression values (*CDKN1A*/*TP53*) was 3.6. With less than 100 million reads, the ratio of the expression levels of the two genes ranged from 2.9 to 15. This ratio fluctuated by as much as 300% at read depths of less than 100 million. However, with more than 100 million reads, the expression ratio ranged from 3.6 to 3.7, and the largest deviation from the ratio obtained was 4%. Thus, we conclude that deep sequencing is necessary to ensure the accurate quantification of relative gene expression.

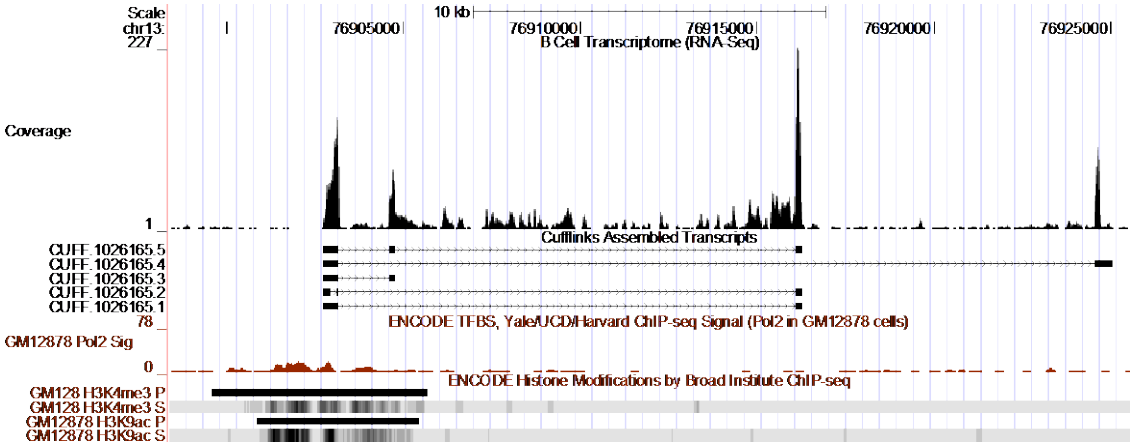
Figure 2.12 Expression levels of *BRD4* versus sequencing depth. FPKM values of four alternatively-spliced transcripts of *BRD4* are shown. With less than 100 million reads, the expression level of *BRD4*-201 (orange line) is overestimated, while that of *BRD4*-204 (purple line) is underestimated. Error bars represent 95% confidence intervals reported by Cufflinks.



2.3.7 Discovery of novel gene models by RNA-Seq

A feature of RNA-Seq is its ability to detect unknown transcripts. As such, we used Cufflinks without known gene models to annotate transcribed fragments and identified 230,006 genes. The majority (77%) of these have already been identified by gene annotation groups such as Aceview, CCDS, Gencode, Mammalian Gene Collection, RefSeq, UCSC, and Vega. Of the remaining 53,939 transcribed fragments, 6,892 (13%) overlap RNA polymerase II binding sites (Rosenbloom et al. 2010). After filtering out known genes and fragments that overlap repetitive genomic regions (Self Chain and RepeatMasker tracks on the UCSC Genome Browser), we have 801 “unknown” genes. These genes have relative high expression (mean FPKM = 95), but they are quite short; the average length of these “unknown” genes is 0.9 kb compared to 1.8 kb for known genes. Furthermore, only 21 of these novel genes have alternatively-spliced transcripts; we show as example a multi-isoform gene on chromosome13:76902733-76925064 that has five alternatively-spliced transcripts and an FPKM value of 1400 (Figure 2.13). Support for the validity of this “unknown” gene includes an upstream 59 RNA polymerase II peak and overlaps with histone H3K4Me3 and H3K9Ac marks. These findings suggest that with our data set of about 40 million reads per sample, we detected most of the known genes (polyadenylated mRNAs) in B-cells.

Figure 2.13 Newly identified gene on chromosome 13. This gene has five alternatively-spliced transcripts. The RNA polymerase II peak and H3K4Me3 andH3K9Ac marks are located at the 5' ends of the gene.



2.4 Discussion

In this chapter, we obtained 879 million 50-bp RNA-Seq reads derived from cultured B-cells of 20 CEPH individuals to characterize the human B-cell transcriptome and to determine the coverage needed for various RNA-Seq studies. We mapped 80% of our sequence reads to the human reference genome, of which 84% aligned to unique locations. We found that with 100 million reads, the number of aligned reads increased with sequencing depth; however, with read depths greater than 100 million, the percentages remained constant. In contrast, the percentage of aligned reads that map unambiguously to the genome was constant at 84% for all sequencing depths.

We detected 20,776 Gencode genes and 67,453 of their alternatively-spliced transcripts using an FPKM threshold of 0.05. More than 90% of multi-exon genes are alternatively-spliced, but their isoforms are not expressed at similar levels. Rather, the majority of genes have one isoform that is expressed at higher levels than the other isoforms. In our expression analysis, we used an FPKM cutoff for expression because inclusion of all transcripts with FPKM values greater than zero will include some very small FPKM measurements. We accompany the use of this threshold with two caveats. First, this threshold is just a means of evaluation and should not be taken to define gene expression. There are transcripts with FPKM values less than 0.05 that are, indeed, expressed. Secondly, our results suggest that the distribution of FPKM values for genes and transcripts varies with respect to sequencing depth (Toung et al. 2011); therefore, the threshold of 0.05 should be considered concurrently with the fact that it was determined using a sequencing depth of 879 million reads.

We assumed that our 879-million pooled data set provides a comprehensive collection of expressed genes and transcripts and their expression levels. We found that with 100 million reads, we detected the majority of genes (81%) and transcripts (90%), but their expression levels were not sufficiently accurate. At 100 million reads, only 6% of genes have FPKM measurements that are within 10% of their "final" values compared to 72% at 500 million reads. Thus deep sequence coverage is needed for gene expression studies. The coverage that we report here probably represents an upper bound of the required depth since the increasing length of sequence reads and the use of paired-end reads will allow more sequences to be mapped, thus reducing the numbers of reads needed to obtain robust expression values.

An enticing feature of RNA-Seq lies in its power to detect transcripts independent of existing information. In this study, we uncovered 801 potential "unknown" genes. Most of these transcribed fragments are short and comprise single exons; only 21 of these genes are alternatively-spliced. While these results do not necessarily undermine the ability to uncover unknown transcripts using RNA-seq, they suggest that with about 40 million reads per sample, we can detect most of the known genes in human B-cells. In summary, recent advances in sequencing technologies have allowed us to obtain deep coverage of human B-cell transcriptomes at single-nucleotide resolution. Our results provide some guidelines for the design of gene expression studies. The B-cells in this study have been used in many other functional (Linsley et al. 1991; Peters et al. 1991; Stern et al. 1984) and genetic studies (Dixon et al. 2007; Dolan et al. 2004; Morley et al. 2004); detailed information on gene expression and structure will extend the previous

analyses and facilitate future projects. Our data are available as the “B-Cell Transcriptome (RNA-seq)” track on the UCSC Genome Browser.

2.5 Materials and Methods

2.5.1 Samples

Immortalized lymphoblastoid cell lines for 20 European-derived individuals from the Utah pedigrees of the Center d'Étude du Polymorphisme Humain collection (CEPH) were obtained from Coriell Cell Repositories. No individuals were known to be blood relatives, and there is no known history of major medical illness. Specifically, the individuals (10 males and 10 females) are GM06985, GM07000, GM07034, GM07055, GM07056, GM07345, GM11832, GM11839, GM11992, GM11993, GM11994, GM12056, GM12145, GM12155, GM12716, GM12717, GM12750, GM12813, GM12872, and GM12891.

Cells were grown to a density of 5×10^5 cells/mL in RPMI 1640 supplemented with 15% fetal bovine serum, 100 units/mL penicillin-streptomycin, and 2 mM L-glutamine. Cells were harvested 24 hours after addition of fresh medium. Total RNA was extracted from cell pellets using the RNeasy Mini-Kit with DNase treatment (Qiagen).

2.5.2 RNA-Sequencing

RNA-Seq was performed as recommended by the manufacturer (Illumina). Briefly, poly(A) mRNA was fragmented, and first-strand cDNA was prepared using random hexamers. Following second-strand cDNA synthesis, end repair, addition of a single A base, adaptor ligation, agarose gel isolation of ~200-bp cDNA, and PCR amplification, the samples were sequenced using the Illumina 1G Genome Analyzer.

2.5.3 Alignment and isoform abundance estimation

Sequence reads were mapped using TopHat (v. 1.1.4) (Trapnell et al. 2009) with default settings. Data sets larger than 300 million reads were randomly split into equal

subsets ranging from two to four because of memory limitations. Cufflinks (v. 0.9.3) (Trapnell et al. 2010) was then used to assemble reads into transcripts and estimate their abundances. Cufflinks was run (1) with the Gencode reference annotation (v. 3c) (Harrow et al. 2006) to generate FPKM values for known gene models and (2) without an annotation file to create gene bundles representing potential novel transcribed fragments.

2.5.4 Sampling selection of sequence reads

Across the 20 samples, 43,819,745 ± 8,194,875 (mean ± standard deviation) reads were obtained. All reads from each sample were pooled to form a data set consisting of 878,668,290 reads (879-million data set). To investigate the effect of sequencing depth on RNA-Seq data, we randomly selected reads from the pooled data set and created subsets varying from 1 to 9 million reads total (in intervals of 1 million reads), 20 to 90 million reads total (in intervals of 10 million reads), and 100 to 700 million reads total (in intervals of 100 million reads). To ensure that the particular reads selected for each sequencing depth are fairly representative, we randomly sampled 100 million reads from the pooled 879-million data set 10 times and analyzed the overall alignment statistics obtained across the 10 random samplings. The percentage of total reads aligning to the genome across the 10 randomizations was 80% ± 0.005% (mean ± standard deviation), of which 84% ± 0.005% (mean ± standard deviation) aligned to unique locations. Overall, the alignment statistics are similar across the 10 random samplings, indicating that the particular reads chosen in each of our sample sizes is representative. Furthermore, using Cufflinks, we carried out analyses to ensure that expression values are not affected by the samplings of reads. We found that across the 10 samplings, 17,967 genes and 69,672 transcripts were detected. Eighty-eight percent of the genes and transcripts were detected

in all samplings. The coefficients of variation of the FPKM values for these genes across the 10 data sets were 0.10 ± 0.16 at the gene-level and 0.49 ± 0.53 at the transcript-level (mean ± standard deviation). Thus, the expression levels of genes and transcripts in different samplings of 100 million reads are fairly stable.

2.5.5 RNA-Seq and microarray analyses

For all analyses in which RNA-Seq data were compared with microarray data previously generated (GSE12526), RNA-Seq data were $\log_2$-normalized. Prior to $\log_2$ transformation, we added 2 to the FPKM values to avoid negative values after the $\log_2$ transformation.

Chapter 3. RNA-DNA Sequence Differences in Humans

3.1 Abstract

The transmission of sequence information from DNA to RNA is a critical process. In this chapter, we compared RNA sequences from human B-cells of 27 individuals to the corresponding DNA sequences from the same individuals and uncovered more than 10,000 exonic sites where the RNA sequences do not match that of the DNA. All 12 possible categories of discordances were observed. These differences were nonrandom as many sites were found in multiple individuals and in different cell types including primary skin cells and brain tissues. Using mass spectrometry, we detected peptides that are translated from the discordant RNA sequences and thus do not correspond exactly to the DNA sequences. These widespread RNA-DNA differences in the human transcriptome provide a yet unexplored aspect of genome variation.

This chapter contains results from a previously published report (M. Li et al. 2011) and includes work from members of the Vivian Cheung laboratory.

3.2 Introduction

DNA encodes genetic information that is transmitted into messenger RNA (mRNA) and proteins that perform cellular functions. It is assumed that the sequence of mRNA accurately reflects that of DNA because mRNA serves as the template for synthesis of proteins. As such, genetic studies have mainly focused on individual variation in DNA sequence as the basis of differences in disease susceptibility. Prior to advances in sequencing technology that allow for comprehensive analysis of RNA sequence information, studies of mRNA focused mainly on variation in expression levels and not sequence differences among individuals.

There are, however, known mechanisms that generate differences between the sequence of mRNA and that of the underlying genomic template. These processes include transcriptional infidelity (Libby & Gallant 1991; Sydow & Cramer 2009) and RNA editing (Bass & Weintraub 1988; S. H. Chen et al. 1987; Powell et al. 1987). Transcriptional errors are rare as proofreading and repair mechanisms exist to ensure high accuracy in RNA synthesis (M. J. Thomas et al. 1998; D. Wang et al. 2009; Zenkin et al. 2006). RNA editing, or site-specific sequence alteration of RNA transcripts, is performed by enzymes that target mRNA posttranscriptionally – ADARs (adenosine deaminases that act on RNA) mediate A-to-I (or A-to-G) changes by deamination of adenosine to inosine, which is then recognized by the translation machineries as a guanosine and APOBECs (apolipoprotein B mRNA editing enzymes, catalytic polypeptide-like) cause C-to-U changes by deamination of cytidine to uridine. Previously, sequence comparisons and computational predictions have identified many A-to-G editing sites (Athanasiadis et

al. 2004; Levanon et al. 2004; J. B. Li et al. 2009; Sakurai et al. 2010). In contrast, C-to-U changes are rare, as apolipoprotein B is one of the few known target genes of human APOBEC1 (Chester et al. 2000; Conticello 2008).

We obtained sequences of DNA and RNA samples from immortalized B cells of 27 unrelated Centre d'Etude du Polymorphisme Humain (CEPH) (Dausset et al. 1990) individuals, who are part of the International HapMap (International HapMap Consortium 2003, 2005) and the 1000 Genomes (1000 Genomes Project Consortium 2010) projects. When we compared the DNA and RNA sequences of the same individuals, we found 28,766 events at over 10,000 exonic sites that differ between the RNA and the corresponding DNA sequences. Each of these differences was observed in at least two individuals, and many were seen in primary skin cells and brain tissues from a separate set of individuals and in expressed sequence tags (ESTs) from cDNA libraries of various cell types. About 43% of the differences are transversions and therefore cannot be the result of typical deaminase-mediated RNA editing. By mass spectrometry, we also found peptide sequences that correspond to the RNA variant sequences, but not the DNA sequences, suggesting that the RNA forms are translated into proteins.

3.3 Results

3.3.1 RNA and DNA samples

We compared the DNA and RNA sequences from B-cells of 27 unrelated CEPH individuals (Table 3.1). We chose these samples because much information is available on them, including dense DNA genotypes obtained using different technologies (Cann 1992; Matise et al. 2003). The genomes of B-cells from the CEPH collection are stable as evidenced by Mendelian inheritance of genetic loci that allowed the construction of microsatellite- and single-nucleotide polymorphism (SNP)-based human genetic maps (Cann 1992; Matise et al. 2003). More recently, the International HapMap Consortium (International HapMap Consortium 2003, 2005) obtained millions of SNP genotypes, and the 1000 Genomes Project (1000 Genomes Project Consortium 2010) sequenced the DNA of these individuals. Comparison of genotype data from these two projects showed high concordance (~99%). For our analyses of RNA-DNA sequence differences, we used the DNA genotypes and sequences from the two projects and considered the following two types of sites:

First, we considered sites that are monomorphic in the human genome. We defined a monomorphic site as one in which there is no evidence for sequence variation at that locus in dbSNP, the HapMap, and the 1000 Genomes Project. Different studies have analyzed these 27 and hundreds of additional individuals for DNA variants, and thus, if a site has not been identified as polymorphic, most likely all individuals have the same genotypes at these sites. However, in order to be certain of the genotypes at monomorphic sites in these 27 individuals, we compared their DNA sequences from the 1000 Genomes project with the sequences of the human reference genome by traditional

Sanger sequencing (Sanger et al. 1977). To be included in our comparison analysis, we required that each site be covered by at least four reads in the 1000 Genomes Project and that the sequences from 1000 Genomes be the same as those of the reference genome. To ensure the integrity of the aliquots of B-cells that we used for analyses, we carried out Sanger sequencing of their DNA and found perfect concordance of sequences with data from the 1000 Genomes (thus also the reference genome sequences) (Table 3.2).

Second, we considered SNP sites. For each individual, a SNP locus was included only if HapMap as well as the 1000 Genomes Project reported the same homozygous genotype. We have high confidence in these sites because despite using different technologies (microarray-based genotyping in HapMap and high-throughput sequencing in 1000 Genomes), we obtained identical genotypes in the two projects.

We sequenced the RNA of B-cells from the 27 CEPH individuals using high-throughput sequencing technology from Illumina (Bentley et al. 2008). The resulting RNA-Seq reads were mapped to Gencode genes (Harrow et al. 2006) in the reference human genome. In total, we generated approximately 1.1 billion reads of 50 base-pairs (bp) (roughly 41 million reads and 2 Gigabases of sequence per individual), of which about 69% mapped uniquely to the transcriptome (see Materials and Methods). To be confident of the base calls, for each individual, we focused our analysis on high-quality reads (Phred quality score ≥ 25) and sites that were covered by at least 10 uniquely mapped reads. We compared our sequences from these sites to RNA-Sequencing data from another recent study (Montgomery et al. 2010) on the same individuals sequenced to lower depths and found the concordance rate to be greater than 99.5%. This is reassuring given that the samples were prepared and sequenced in different laboratories.

Table 3.1 Statistics on RNA-Sequencing and RNA-DNA sequence differences. Statistics on the RNA-Seq data and sequence differences between RNA and DNA for the 27 individuals are provided.

Sample ID	Sex	Total reads sequenced	Reads aligned uniquely	Total bases at sites with coverage ≥ 10	DNA coverage	# of RDDs at homozygous SNP sites	# of RDDs at monomorphic sites	# of total RDDs
GM06985	F	38,424,688	24,864,586	16,038,399	2.6	14	394	408
GM06994	M	37,634,150	31,439,915	19,979,779	4.6	36	1208	1244
GM07000	F	36,665,036	26,987,933	16,357,700	4.1	45	1512	1557
GM11829	M	37,293,182	28,117,279	16,305,573	6	17	1011	1028
GM11830	F	36,412,185	26,729,286	16,388,938	6.4	31	938	969
GM11831	M	38,923,882	26,175,492	16,683,852	6.4	29	965	994
GM11832	F	38,657,782	27,779,737	16,629,265	2.8	27	956	983
GM11881	M	49,771,575	29,645,009	15,899,948	5.1	29	1462	1491
GM11992	M	63,260,299	41,104,692	17,703,651	5.7	53	1810	1863
GM11993	F	38,441,615	27,190,601	15,681,535	3.7	12	621	633
GM11994	M	42,390,584	27,741,424	16,702,144	13.8	18	466	484
GM12003	M	36,977,676	29,073,057	17,727,318	5.9	46	1384	1430
GM12004	F	28,390,919	21,836,374	13,369,507	6.2	6	480	486
GM12005	M	45,286,259	25,260,540	11,909,776	6.1	20	1248	1268
GM12006	F	35,506,222	28,101,550	14,554,102	3.3	34	1070	1104
GM12043	M	36,999,366	27,453,229	17,312,508	4.2	26	728	754
GM12044	F	39,839,492	30,992,207	16,088,150	4.7	40	1255	1295
GM12144	M	35,020,229	21,561,396	11,631,101	5.5	45	1334	1379
GM12155	M	43,551,373	33,153,306	16,821,880	11.7	9	273	282
GM12716	M	46,656,852	30,454,296	16,024,579	5.9	45	1358	1403
GM12717	F	40,635,780	32,974,555	20,054,384	5.2	25	1339	1364
GM12750	M	45,066,631	31,480,502	17,290,976	4.2	32	1319	1351
GM12762	M	47,651,198	31,766,092	17,761,776	2.2	22	819	841
GM12813	F	45,288,662	34,565,796	18,137,367	3.7	35	1179	1214
GM12814	M	39,024,182	29,109,410	16,700,518	4.7	28	996	1024
GM12872	M	45,367,154	17,499,200	9,290,271	4.9	42	927	969
GM12874	M	39,588,081	20,781,628	10,406,285	4.9	29	919	948
Total		1,108,725,054	763,839,092	429,451,282	144.5	795	27971	28766
Mean		41,063,891	28,290,337	15,905,603	5.4	29	1036	1065

Table 3.2 Genotypes at monomorphic sites verified by Sanger sequencing. We verified the genotypes at monomorphic sites using DNA extracted from the pooled samples of cell extracts used in the mass spectrometry analysis. Sanger sequencing verified that all the sites are monomorphic. The results agree with the sequencing data from the 1000 Genomes and Human Genome Projects (as shown in the UCSC Genome Browser for the reference genome).

Gene	RDD	Location (hg18)	genomic DNA (Sanger)
CD22	T-to-G	chr19:40514815	T
DFNA5	T-to-A	chr7:24705225	T
ENO1	T-to-C	chr1:8848125	T
FH	T-to-A	chr1:239747217	T
HMGB1	T-to-A	chr13:29935772	T
HMGB1	A-to-C	chr13:29935469	A
ITPR3	A-to-C	chr6:33755773	A
RAD50	T-to-G	chr5:131979610	T
ROD1	G-to-T	chr9:114026264	G
RPL32	G-to-T	chr3:12852658	G
RPS25P8	A-to-G	chr11:118393375	A
RPS3AP47	C-to-A	chr4:152243651	C
SUPT5H	G-to-T	chr19:44655806	G
TOR1AIP1	T-to-C	chr1:178144365	T
TUBA1	A-to-G	chr2:219823379	A

3.3.2 RNA-DNA sequence differences observed

For each of the 27 individuals, we compared the mRNA sequences from B cells with the corresponding DNA sequences (Figure 3.1). The comparison revealed many sites where the mRNA sequences differ from the corresponding DNA sequences of the same individual. To minimize the chance that these differences are due to sequencing errors, we required that at least 10% of the reads covering a site differ from the DNA sequence and that at least two individuals show the same RNA-DNA difference at the site. We call each occurrence of a difference between RNA and DNA sequences an "event" and the chromosomal location where such a difference occurs a "site". Each person can contribute an event to the site, and thus, there could be multiple events at a site. Among our 27 subjects, we identified 28,766 events where the RNA sequences do not match those of the corresponding DNA sequences. These events are found in 10,210 exonic sites in the human genome and reside in 4741 known genes (36% of 13,214 genes that are covered by 10 or more RNA-Seq reads in at least one part of the gene and in two or more individuals). Using gene orientation information provided by Gencode, we observed all 12 possible categories of base differences between RNA and the corresponding DNA (Figure 3.2). All 12 types of differences were found in each of the 27 samples; the relative proportion of each type is similar across individuals. There are 6,698 A-to-G events, which can be the result of deamination by ADAR. There are 1,220 C-to-T differences, which can also be mediated by a deaminase. However, it is notable that APOBEC1 and its complementation factor A1CF are not expressed in our B-cells [fragments per kilobase of exon per million fragments mapped (FPKM) (Trapnell et al. 2010) ~ 0 for both genes]. As such, it is likely that an unknown deaminase or other

mechanism is involved. Even for relatively well-characterized proteins such as APOBEC1, a recent RNA-Seq study of *Apobec1*$^{-/-}$ mice uncovered many previously unknown targets (Rosenberg et al. 2011). In addition, we found 12,507 transversions (43%), which cannot result from classic deaminase-mediated editing. Because we do not know the mechanism by which these differences between RNA and DNA sequences arise, we refer to them as RNA-DNA differences (RDDs). An example of an RDD is a C-to-A difference on chromosome 12 (at position 54,841,626 bp) in the myosin light chain gene *MYL6*, where 16 of our subjects have C/C in their DNA but A/C in their RNA sequences. Another example is an A-to-C difference on chromosome 6 (at position 44,328,823 bp) in the gene *HSP90AB1* that encodes a heat shock protein, where eight individuals have homozygous A/A DNA genotype but have A/C in their RNA. Additional examples are shown in Table 3.3. These sites where RNA sequences differ from the corresponding DNA sequences appear to be nonrandom because the identical differences were found in multiple individuals: 8163 (80.0%) of the sites were found in at least 50% of the informative individuals (i.e. with RNA-Seq coverage ≥ 10 and DNA-Seq coverage ≥ 4 at the site). Some sites were found in all or nearly all informative individuals. For example, the DNA sequences of all 19 informative individuals at position 49,369,615 bp of chromosome 3 in the GPX1 gene are G/G, whereas their RNA sequences are G/A. The remaining individuals were not included because available data did not meet our inclusion criteria although the data suggest the same RDD in all remaining individuals: G/G in DNA, and G/A in RNA.

Figure 3.1 Workflow for the identification of RNA-DNA sequence differences.

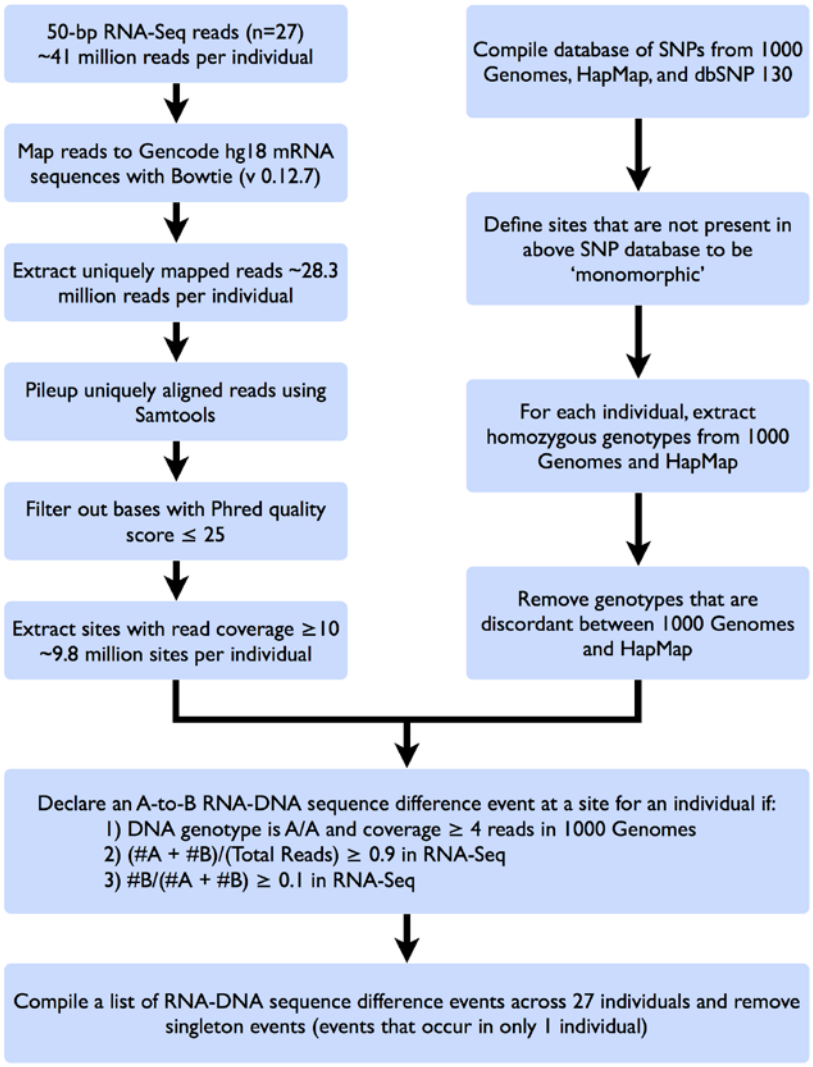


Figure 3.2 Distribution of RNA-DNA sequence difference events across 27 individuals. The number of RNA-DNA sequence difference events detected in human B-cells of 27 individuals is shown below.

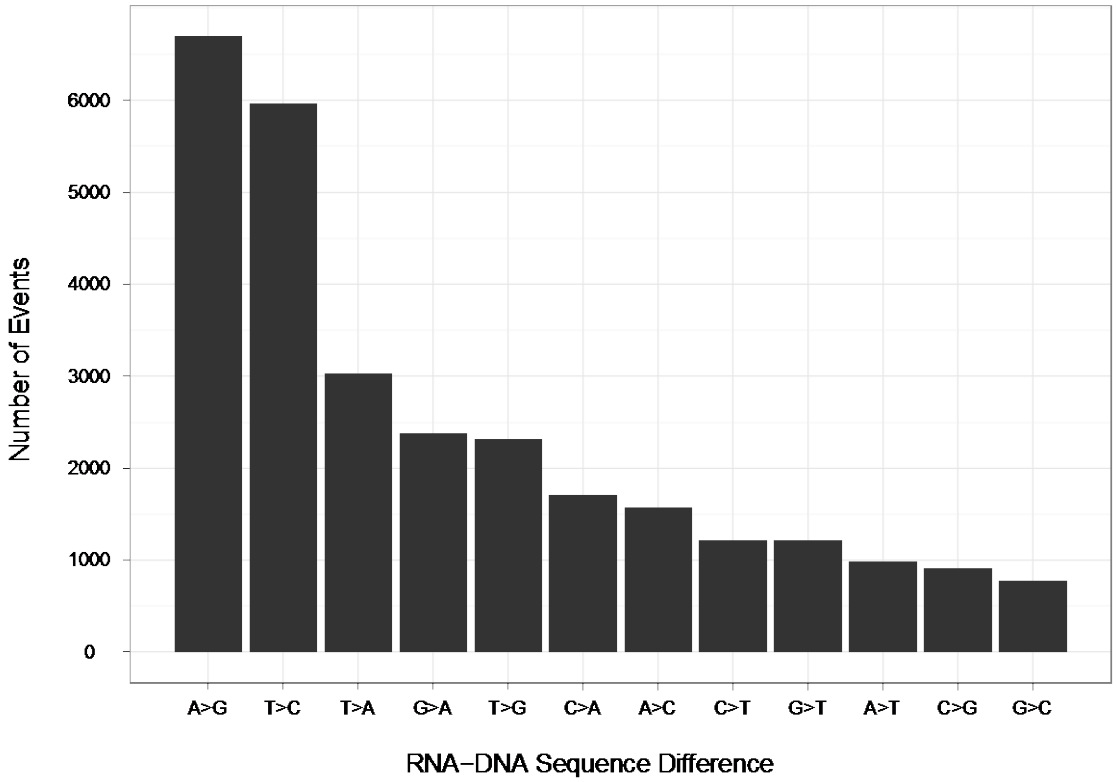


Table 3.3 Selected examples of sites that show RNA-DNA sequence differences in B-cells and EST clones.

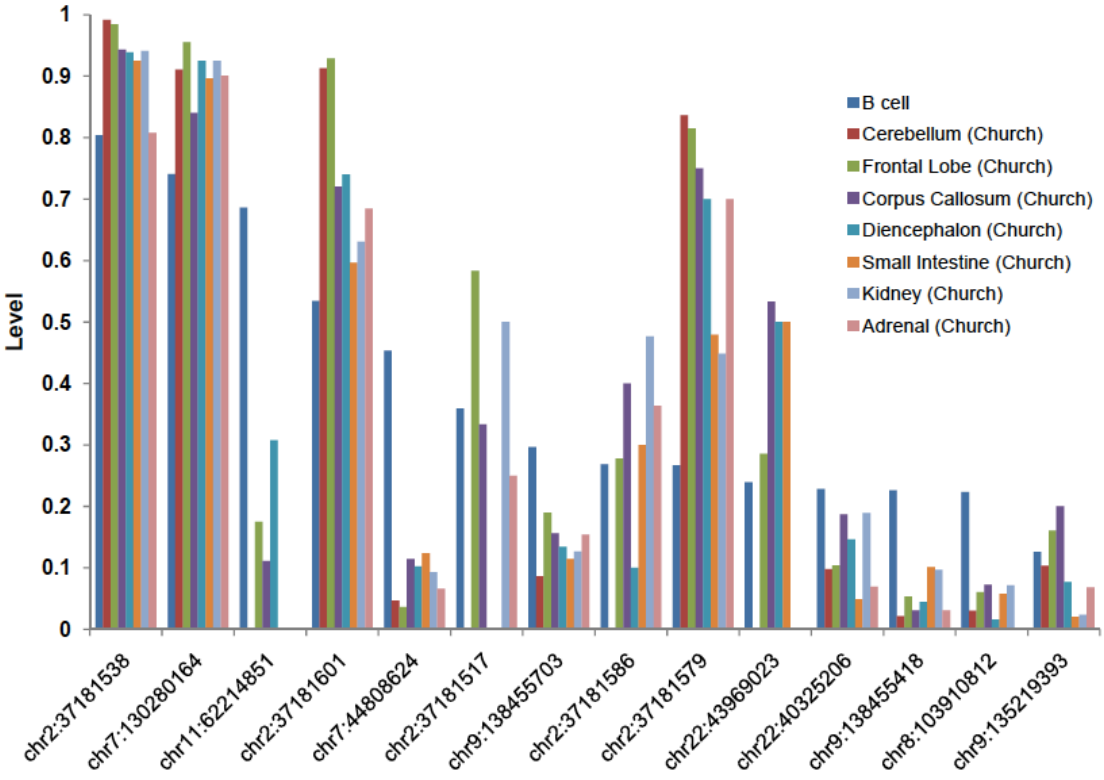
Gene	Chr.	Position (bp)*	Type	Number of informative individuals †‡	Number of individuals with RDD ‡	Average level ‡§	Range ‡§	EST
HSP90AB1	6	44,328,823	A-to-C	11	8	0.39	[0.15, 0.79]	BQ355193 (head, neck), BX413896 (B-cell)
AZIN1	8	103,910,812	A-to-G	17	10	0.22	[0.12, 0.37]	CD359333 (testis), BF475970 (prostate)
CNBP	3	130,372,812	A-to-T	18	16	0.13	[0.10, 0.21]	EL955109 (eye), BJ995106 (hepatoblastoma)
MYL6	12	54,841,626	C-to-A	16	16	0.35	[0.12, 0.60]	EC496428 (prostate), BG030232 (breast adenocarcinoma)
RBM23	14	22,440,217	C-to-G	11	5	0.18	[0.11, 0.35]	BQ232763 (testis, embryonic)
RPL23	17	34,263,515	C-to-T	12	8	0.16	[0.10, 0.22]	BP206252 (smooth muscle), CK128791 (embryonic stem cell)
BLNK	10	97,957,645	G-to-A	14	7	0.14	[0.11, 0.17]	BF972964 (leiomyosarcoma), BE881159 (lung carcinoma)
C17orf70	17	77,117,583	G-to-C	2	2	0.26	[0.24, 0.28]	AA625546 (melanocyte), AA564879 (prostate)
HMGN2	1	26,674,349	G-to-T	7	4	0.22	[0.14, 0.43]	BX388386 (neuroblastoma), BE091398 (breast)
CANX	5	179,090,533	T-to-A	9	8	0.2	[0.13, 0,30]	EL950052, DB558106
EIF3K	19	43,819,430	T-to-C	19	14	0.16	[0.10, 0.27]	AI250201 (ovarian carcinoma), AI345393 (lung carcinoma)
RPL37	5	40,871,072	T-to-G	6	6	0.27	[0.16, 0.45]	CF124792 (T cell), DW459229 (liver)

* hg 18 build of the human genome
† minimum of 10 reads in RNA-Seq and 4 reads in DNA-Seq
‡ in B-cells
§ proportion of reads at site that contain RDD

3.3.3 EST validation of RNA-DNA sequence differences

Computational and experimental validations also upheld the RNA-DNA differences we observed. First, for 120 sites (10 sites per RDD type randomly selected; see Table 3.3 for examples), we looked for evidence of RDD in the human EST database by BLAST alignment (Altschul et al. 1990) and manual inspection of each result. For 81 of the 120 sites, we found EST clones that contain the RDD alleles. The numbers of sites found in human ESTs are similar across different RDD types (average 67.5%; range: 60 to 90%). Second, we examined previously identified A-to-G editing sites (J. B. Li et al. 2009). Fourteen of the A-to-G sites that we identified were found in their data even though different cell types were studied. Even the levels of editing at these sites are similar between the two studies (Figure 3.3). Twelve additional sites were found in both studies but were filtered out because they did not meet our selection criteria.

Figure 3.3 Comparison of A-to-G RNA editing levels in B-cells to those in cell types published by Church and colleagues (J. B. Li et al. 2009).



3.3.4 Sanger sequencing validation of RNA-DNA sequence differences

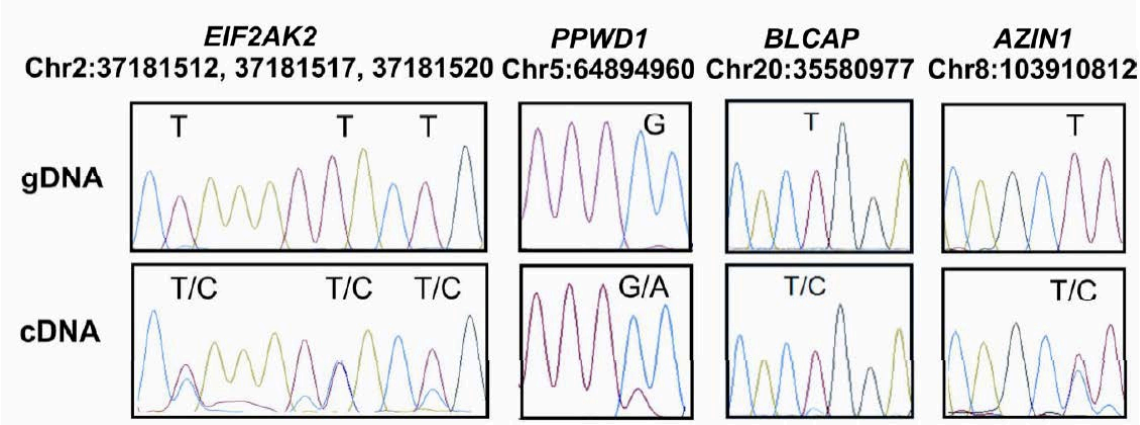
Next, we validated our findings experimentally by Sanger sequencing of both DNA and RNA at 12 randomly selected sites in B-cells (2 to 9 individuals per site), primary skin (foreskin; 8 to 10 individuals per site), and brain cortex (6 to 10 individuals per site). We regrew the B-cells from our subjects and extracted DNA and mRNA from the same aliquots of cells. By sequencing the paired DNA and RNA samples and analysis of each chromatogram by two individuals independently, we confirmed 57 events in 11 sites (Table 3.4 and Figure 3.4). In *EIF2AK2*, in all of the eight individuals whose samples were sequenced, three sites were found within 10 nucleotides (nt) (see Table 3.4). An RNA-DNA sequence difference was not validated in one site in *NDUFC2*. Sanger sequencing is not very sensitive or quantitative; thus, we do not expect to validate all sites, especially those with low levels of RDD. To assess whether RDD shows cell type specificity, we looked for evidence of RNA-DNA sequence differences using primary human cells. We studied the same sites as above by Sanger sequencing of DNA and RNA samples from primary skin fibroblasts and brain (cortex) of a separate set of normal individuals (for each site, we examined the DNA and RNA of 6 to 10 samples per cell type). We identified 55 RDD events in primary skin cells and 62 events in brain cortex (Table 3.4). The results suggest that most sites are shared across cell types (Table 3.4). However there are exceptions, as an A-to-G difference in *EIF2AK2* (chromosome 2: 37,181,512) was only found in B-cells and brain cortex but not in primary skin cells. We also queried the EST database for evidence of RDD (Table 3.3). The RNA alleles are seen in a wide range of tissues from embryonic stem cells to brain and testis; they are also found in tumors such as lung carcinoma and neuroblastoma.

Table 3.4 Sanger sequencing validation of RNA-DNA sequence difference sites.

Gene	Chr.	Position (bp)*	Type	Location	Amino Acid Change	B-cells†		Primary skin fibroblast†		Brain (cortex)†	
						# of Informative Individuals	# of Individuals with RDD	# of Informative Individuals	# of Individuals with RDD	# of Informative Individuals	# of Individuals with RDD
EIF2AK2	2	37,181,512	A-to-G	3′ UTR	Not applicable	8	8	8	0	10	10
	2	37,181,517	A-to-G	3′ UTR	Not applicable	8	8	8	3	10	10
	2	37,181,520	A-to-G	3′ UTR	Not applicable	8	8	8	3	10	10
	2	37,181,538	A-to-G	3′ UTR	Not applicable	8	8	8	6	10	10
AZIN1‡	8	103,910,812	A-to-G	Coding, exonic	S-to-G	2	2	10	0	9	8
DPP7	9	139,128,755	C-to-T	Coding, exonic	Synonymous (P)	9	2	8	1	10	0
PPWD1	5	64,894,960	G-to-A	Coding, exonic	E-to-K	2	2	8	8	8	8
HLA-DQB2	6	32,833,537	G-to-A	Coding exonic	G-to-S	2	2	10	10	NE§	NE
	6	32,833,545	G-to-A	Coding, exonic	R-to-H	2	2	10	10	NE	NE
	6	32,833,550	C-to-T	Coding, exonic	Synonymous (I)	2	2	10	10	NE	NE
BLCAP#	20	35,580,977	A-to-G	Coding, exonic	Q-to-R	6	4	10	4	6	6
NDUFC2	11	77,468,303	C-to-G	Coding, exonic	L-to-V	10	0	10	0	10	0

* hg 18 build of the human genome
† in all cases, matched DNA and RNA from the same individuals were sequenced
‡ also reported by Church and colleagues (J. B. Li et al. 2009)
§ NE, not expressed
# known site that was used as a positive control

Figure 3.4 Examples of Sanger sequencing validation of RNA-DNA sequence difference sites. Sequences of matched DNA and RNA sequences from cultured B-cells for four genes are shown. Each matched pair of DNA and RNA sequences was from the same individual. The reverse strand of *EIF2AK2*, *AZIN1* and *BLCAP* was sequenced. Locations of the sites are based on hg18 build of the human genome.



3.3.5 Proteomic evidence for RNA-DNA sequence differences

Validation at the sequence level is important but does not address concerns such as the difficulty in aligning sequences that are highly similar and errors introduced by enzymes in reverse transcription steps. We believe that such artifacts are unlikely considering the consistent patterns across sequencing methods. An alternative and independent validation would be to ask whether the RNA variants in RDD sites are translated to proteins. To do so, we first searched mass spectrometry data from human ovarian cancer cells (Sodek et al. 2008) and leukemic cells for putative RDD sites. Because the levels of most RDDs are less than 100%, both DNA and the RDD forms of the mRNAs should be available to be translated (hereafter, we refer to mRNAs that correspond identically to the DNA sequences as DNA forms and those that contain an RDD as RNA forms). In the ovarian cancer and leukemic cells, we indeed found examples of proteins with peptides encoded by both DNA and RNA forms of mRNA (Table 3.5). Encouraged by the search results and cognizant of possible genome instability and thus DNA mutations in cancer cells, we carried out mass spectrometry analysis of our B-cells. We analyzed the proteome of our B-cells using liquid chromatography-tandem mass spectrometry and detected peptides for 3,217 proteins. Despite advances in mass spectrometry, far less than 50% of peptides can be detected in most studies (de Godoy et al. 2006; Michalski et al. 2011). With a false discovery rate (FDR) less than 1%, we identified 327 peptides that cover RDD sites: 299 and 28 of them correspond to the DNA and RNA forms respectively. For 17 RDD sites, peptides that correspond to both DNA and RNA forms were identified (Table 3.6). By BLAST alignment, we ensured that these 28 peptides were unique to the genes that contain the

RDD sites. In addition, we sequenced the DNA of the B-cells used for mass spectrometry and validated that the DNA sequences were the same as those of the reference genome but differed from the RNA sequences and thus did not encode the RNA forms of the peptides (Table 3.2). It is easier to detect more abundant proteins by mass spectrometry; for most RDD sites, the unaltered DNA forms are more abundant than variant RNA forms of mRNA (Figure 3.5). Thus, it is not surprising to find more peptides that correspond to the DNA than to the RNA sequences. However, the counts of peptides corresponding to the DNA and RNA forms of RDDs should not be taken as a measure of the proportions of DNA versus RNA forms of mRNA that are translated because differences in the amino acid sequences of the peptides affect the ability of mass spectrometry to detect them. In addition, the inability to detect a peptide does not signify its absence from the sample, but may instead be a function of sampling variation. The proteomic data provide an independent validation that mRNA sequences are not always identical to DNA sequences and demonstrate that RNA forms of genes are translated to proteins. They also show that there are peptides in human cells that are not exactly encoded by the DNA sequences. An example of a protein variant that results from RDD is a T-to-A RNA-DNA sequence difference at chr19:60,590,467 in *RPL28* that leads to the loss of a stop codon and gain of 55 amino acids. We identified peptides corresponding to the 55-amino acid extension of RPL28 protein in the ovarian cancer cells and in our B cells (Figure 3.5). Previously identified cases of RNA editing leading to proteins that contain amino acids not encoded by genomic DNA, such as apolipoprotein B (S. H. Chen et al. 1987; Powell et al. 1987), serotonin and glutamate receptors (Burns et al. 1997; Lomeli et al. 1994; Maas et al. 2001) in humans, and plant ribosomal protein S12

(Phreaner et al. 1996), also support our hypothesis that RDD leads to protein isoforms that do not correspond to the DNA sequences of the encoding genes.

Table 3.5 Detection of peptides encoding DNA or RNA forms by mass spectrometry in ovarian cancer and leukemia cells at multiple RDD sites.

Peptides detected from HOC-7 ovarian cancer cell line					
Protein	**Location***	**RDD**	**Amino acid change**	**DNA form of peptide**	**RNA form of peptide**
RPL28	chr19:60590467	T-to-A	STOP-to-R	N/A	SLIGTASEPR
TUBA1	chr2:219823379	A-to-G	E-to-G	EDMAALEK	EDMAALGK
HSPA4	chr5:132459768	T-to-G	V-to-G	TSTVDLPIENQLLWQIDR	TSTGDLPIENQLLWQIDR
HSP90AB1	chr6:44328823	A-to-C	T-to-P	LVSSPCCIVTSTYGWTANMER	LVSSPCCIVTSTYGWPANMER
Peptides detected from K562 leukemia cell line					
Protein	**Location***	**RDD**	**Amino acid change**	**DNA form of peptide**	**RNA form of peptide**
DHX15	chr4:24187092	T-to-A	Y-to-N	YYDILK	NYDILK
ENO3	chr17:4800624	T-to-G	V-to-G	N/D	LAQSNGWGGMVSHR
FABP3	chr1:31618424	T-to-A	W-to-R	N/D	MVDAFLGTR
HNRNPA3	chr2:177789664	T-to-C	S-to-P	N/D	GGGPGNFMGR
HNRNPL	chr19:44021390	C-to-A	Q-to-K	AITHLNNNFMFGQK	AITHLNNNFMFGKK
MYH10	chr17:8357846	T-to-G	W-to-G	N/D	HWQWGR
RANBP2	chr2:108765620	T-to-A	I-to-K	ITMELFSNIVPR	KTMELFSNIVPR
USP34	chr2:61329192	T-to-A	Y-to-N	N/D	LDMTPNTEDFLMGK
ZFP91	chr11:58138282	A-to-G	H-to-R	N/D	YLQHRIK

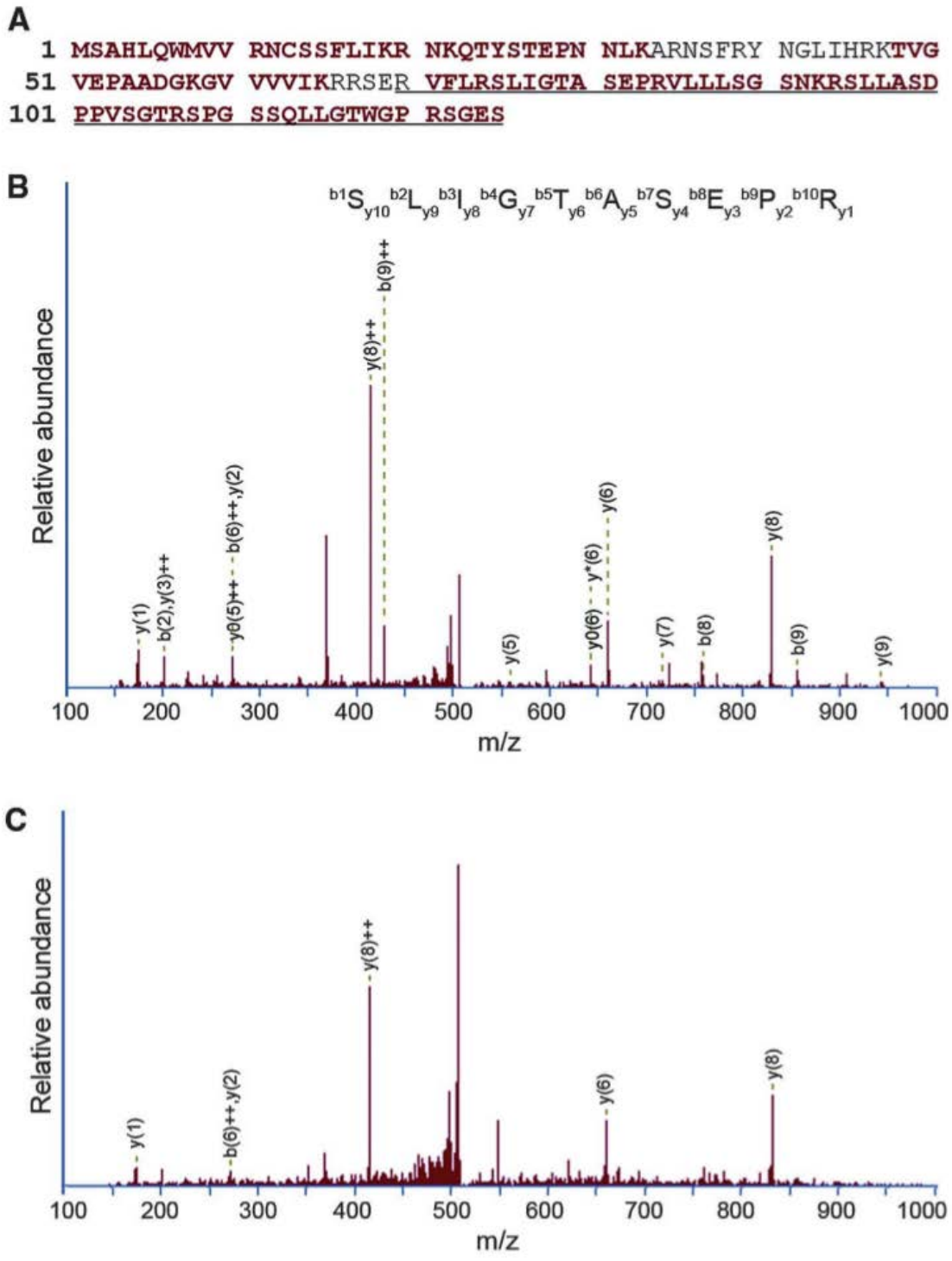
* hg18 build of human genome
N/A: not applicable
N/D: not detected

Table 3.6 Peptides encoded by both the DNA and RNA forms.

Protein	Position*	RDD	Amino acid change	DNA form†	RNA form†
AP2A2	chr11:976858	T-to-G	Y-to-D	YLALESMCTLASSEFSHEAVK	DLALESMCTLASSEFSHEAVK
DFNA5‡	chr7:24705225	T-to-A	L-to-Q	VFPLLLCITLNGLCALGR	VFPQLLCITLNGLCALGR
ENO1	chr1:8848125	T-to-C	L-to-P	EGLELLK	EGPELLK
ENO3	chr17:4800624	T-to-G	V-to-G	LAQSNGWGVMVSHR	LAQSNGWGGMVSHR
FABP3	chr1:31618424	T-to-A	W-to-R	MVDAFLGTWK	MVDAFLGTR
FH‡	chr1:239747217	T-to-A	I-to-K	IEYDTFGELK	KEYDTFGELK
HMGB1	chr13:29935772	T-to-A	Y-to-N	MSSYAFFVQTCR	MSSNAFFVQTCR
NACA	chr12:55392932	G-to-A	D-to-N	DIELVMSQANVSR	NIELVMSQANVSR
NSF	chr17:42161411	T-to-C	V-to-A	LLDYVPIGPR	LLDYAPIGPR
POLR2B	chr4:57567852	T-to-A	L-to-Q	IISDGLK	IISDGQK
RAD50‡	chr5:131979610	T-to-G	L-to-R	WLQDNLTLR	WRQDNLTLR
RPL12	chr9:129250509	A-to-G	N-to-D	HSGNITFDEIVNIAR	HSGDITFDEIVNIAR
RPL32‡	chr3:12852658	G-to-T	A-to-S	AAQLAIR	SAQLAIR
RPS3AP47‡	chr4:152243651	C-to-A	T-to-K	EVQTNDLK	EVQKNDLK
SLC25A17	chr22:39520485	A-to-G	E-to-G	TTHMVLLEIIK	TTHMVLLGIIK
TUBA1‡	chr2:219823379	A-to-G	E-to-G	EDMAALEK	EDMAALGK
TUBB2C	chr9:139257297	G-to-A	G-to-D	LHFFMPGFAPLTSR	LHFFMPDFAPLTSR

* hg18 build of the human genome
† for each peptide, the amino acid that differs between the DNA and RNA forms is underlined
Abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
‡ DNA sequences of these and other proteins were verified by Sanger sequencing (Table 3.2)

Figure 3.5 Identification of peptides encoded by both RNA and DNA forms. (**A**) The RNA form of an RDD leads to the loss of a stop codon in RPL28 and extension of 55 amino acids. Peptides detected by mass spectrometry are shown in red. (**B** and **C**) Tandem mass spectrometry (MS-MS) data confirm the detection of peptides encoded by the RDD-containing *RPL28* mRNA. The representative spectra of one peptide (SLIGTASEPR) from ovarian cancer cells (B) and cultured B-cells (C) are shown.

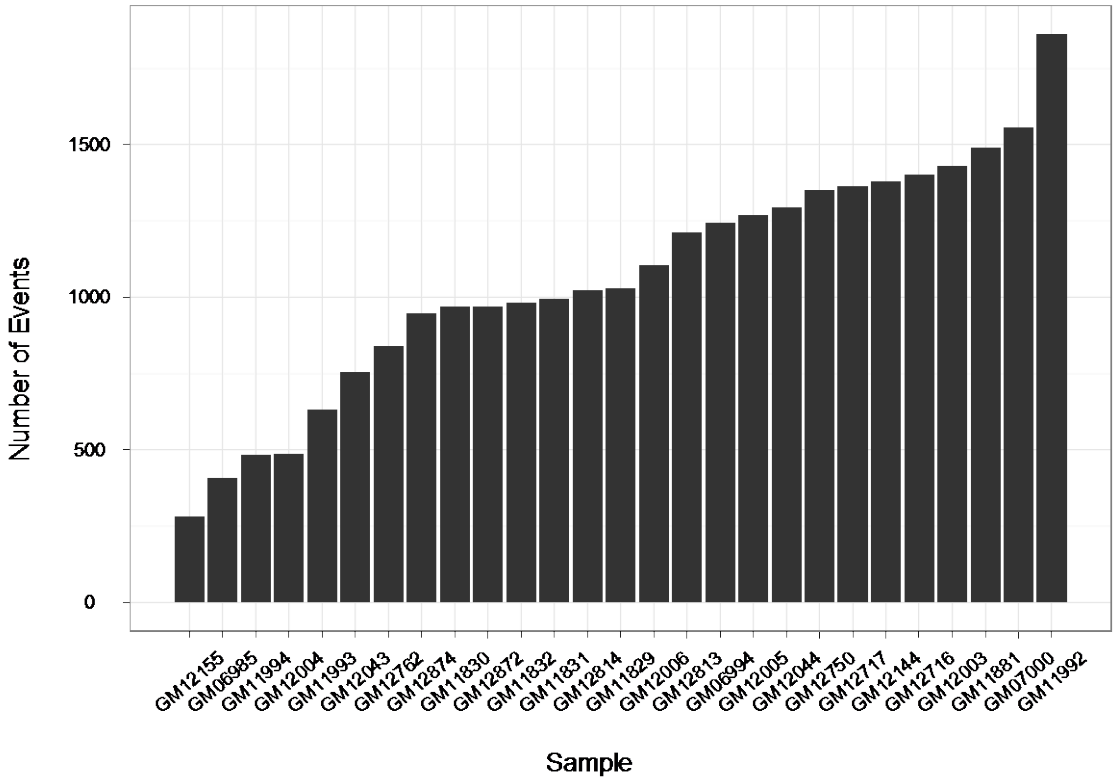


3.3.6 Variation in abundance of RNA-DNA sequence differences

Using our selection criteria, we found that across the 27 individuals, there are 1,065 exonic RDD events on average per sample. The number of events varied among the 27 individuals (range: 282 to 1863) by up to six-fold (Figure 3.6). To evaluate whether the abundance of RDD events is correlated with depth of sequencing, we fitted a linear model using the number of events as the outcome variable and the number of bases covered by 10 or more reads as the independent variable and found that the number of events and sequencing depth are not dependent ($P = 0.51$). Similarly, the depth of DNA sequences from the 1000 Genomes Project cannot explain individual differences in the number of RDD events ($P = 0.12$). From these results, we conclude that we do not have significant evidence to show that variation in RDD frequency is determined by sequencing depth.

We found no significant correlation between *ADAR* expression and the number of RDDs or the numbers of A-to-G events ($P > 0.5$). Thus, we conclude that either *ADAR* expression does not affect the number of editing or RDD events or that our sample size is not sufficient to detect the correlation.

Figure 3.6 Number of RNA-DNA sequence difference events across 27 individuals.



3.3.7 Characteristics of RNA-DNA sequence differences

The 10,210 sites that showed RNA-DNA sequence differences are not evenly distributed across the genome: chromosome 19 has the most sites, whereas chromosome 13 has the fewest. This pattern is observed after correction for differences in size and gene density among chromosomes. RDD sites are significantly ($P < 10^{-10}$) enriched in genes that play a role in helicase activity, and in protein and nucleotide binding (Table 3.7). The 10,210 sites that showed RNA-DNA sequence differences are not evenly distributed within genes. About 44% (4,453 sites) are located in coding exons (10% were found in the last exons), 4% (386 sites) are in the 5' untranslated regions (UTRs), and 39% (3,977 sites) are in the 3' UTRs (Table 3.8; those remaining cannot be classified because of differences in gene structures across isoforms). The results suggest that there are more sites in the 3' ends than in the 5' ends of genes. This pattern was also observed in deamination-mediated RNA editing (Hundley et al. 2008; Rosenberg et al. 2011). Seventy-one percent of the coding sites result in non-synonymous amino acid changes, including 2.1% that lead to the gain or loss of a stop codon in proteins. Relative to other structural features in genes, we found that 4% of RDD sites are within 2 nt of exon borders and 5% are within 30 nt of polyadenylate [poly(A)] signals (Table 3.8). Among RDD types, the numbers of sites near splice junctions are quite similar, but the numbers near poly(A) sites are more different. C-to-A and G-to-A differences are found more often near poly(A) sites. Sites also tended to cluster: 2,613 sites (26%) are within 25 bp of another RDD site, and 1,059 sites (10%) are adjacent to each other. Statistical analysis using a runs test supports the nonrandom locations of the sites (median $P = 0.22$). We did not find obvious patterns or associations with motifs shared across the sites, except for

the A-to-G and A-to-C differences that show a preference for a cytidine 5’ to the adenosine, as previously observed in ADAR-mediated A-to-G changes (Athanasiadis et al. 2004; Maas et al. 2001).

Table 3.7 Most significant gene ontology enrichments for genes containing RDDs.

GO Group (Molecular Function)	Examples	*P*
3'-5' DNA helicase activity	BLM, ERCC3, RECQL, RECQL4, WRN	0
protein binding	ACBD3, AKT2, DENND4A, IL15, MLX	$1x10^{-79}$
catalytic activity	ACLY, CASP3, DHDDS, KDM2A, XRN1	$3x10^{-37}$
RNA binding	APTX, BARD1, EIF1, RBM4, RNASEN	$5x10^{-37}$
nucleotide binding	CDK2, CHD1, DARS, PIF1, U2AF1	$4x10^{-31}$
ATP binding	ASS1, ATP11A, CEHK1, GART, KIF11	$1x10^{-22}$

Table 3.8 Location of RNA-DNA sequence differences within genes.

	5’ UTR		Coding Exon		3’ UTR		Mixed*		Splice		3’ Poly A		Total
	# of sites	%	# of sites	%	# of sites	%	# of sites	%	# of sites	%	# of sites	%	
A- to- C	23	3.8	334	55.5	153	25.4	92	15.3	20	3.3	4	0.7	602
A- to- G	44	1.9	984	42.3	1028	44.2	272	11.7	97	4.2	174	7.5	2328
A- to- T	26	7.2	148	41.2	104	29	81	22.6	17	4.7	7	1.9	359
C- to- A	11	2.1	208	39.8	220	42.1	84	16.1	17	4.7	77	14.7	523
C- to- G	17	5.8	159	53.9	79	26.8	40	13.6	25	8.5	10	3.4	295
C- to- T	56	14.9	178	47.2	78	20.7	65	17.2	3	0.8	4	1.1	377
G- to- A	15	2	342	44.9	292	38.3	113	14.8	46	6	73	9.6	762
G- to- C	27	11.2	114	47.1	56	23.1	45	18.6	14	5.8	2	0.8	242
G- to- T	43	10.8	192	48	110	27.5	55	13.8	24	6	3	0.8	400
T- to- A	16	1.5	398	36.8	550	50.8	118	10.9	35	3.2	90	8.3	1082
T- to- C	58	2.5	953	41.8	951	41.7	316	13.9	33	1.4	142	6.2	2278
T- to- G	50	5.2	443	46	356	37	113	11.7	39	4.1	25	2.6	962

Splice are sites within 2 nt of exon borders
3’ Poly A are sites within 30 nt 3' of AAUAAA sites
* Mixed means in some isoforms the sites are in coding exons and in other isoforms they are in UTRs
Percentages are calculated within each RDD type.

3.3.8 Levels of RNA-DNA sequence differences

We examined the percentage or proporation of mRNAs that differ in sequence from the corresponding DNA at a given site. For each site, to determine the RDD level, we counted the number of reads with a different nucleotide from that in the underlying DNA sequence. The distribution of the level is bimodal (Figure 3.7); the average level is 20% (median = 13%). For some sites, RDD was detected in nearly 100% of the RNA sequences such as the A-to-C difference in the gene that encodes an mRNA decapping enzyme, *DCP1A* (chr3:53,297,343). Sites found in more than 50% of the informative individuals tend to have higher levels of RNA editing or RDD than other sites ($P < 10^{-5}$; Figure 3.8). The levels also differ across individuals. For example, at a G-to-A site in the gene *RHOT1*, which encodes a RAS protein that plays a role in mitochondrial trafficking (chr17:27,526,465), in one individual, the level was 90% while in another sample, it was only 18%. We identified 437 sites with 10 or more informative individuals where the individuals with the highest levels and the lowest levels differed by two-fold or more (range: 2- to 8.6-fold).

Figure 3.7 Distribution of RNA-DNA sequence difference levels. The density of the distribution of RDD levels among the 27 unrelated individuals are shown here. RDD level is measured as the proportion of transcripts at an RDD site bearing the sequence difference allele.

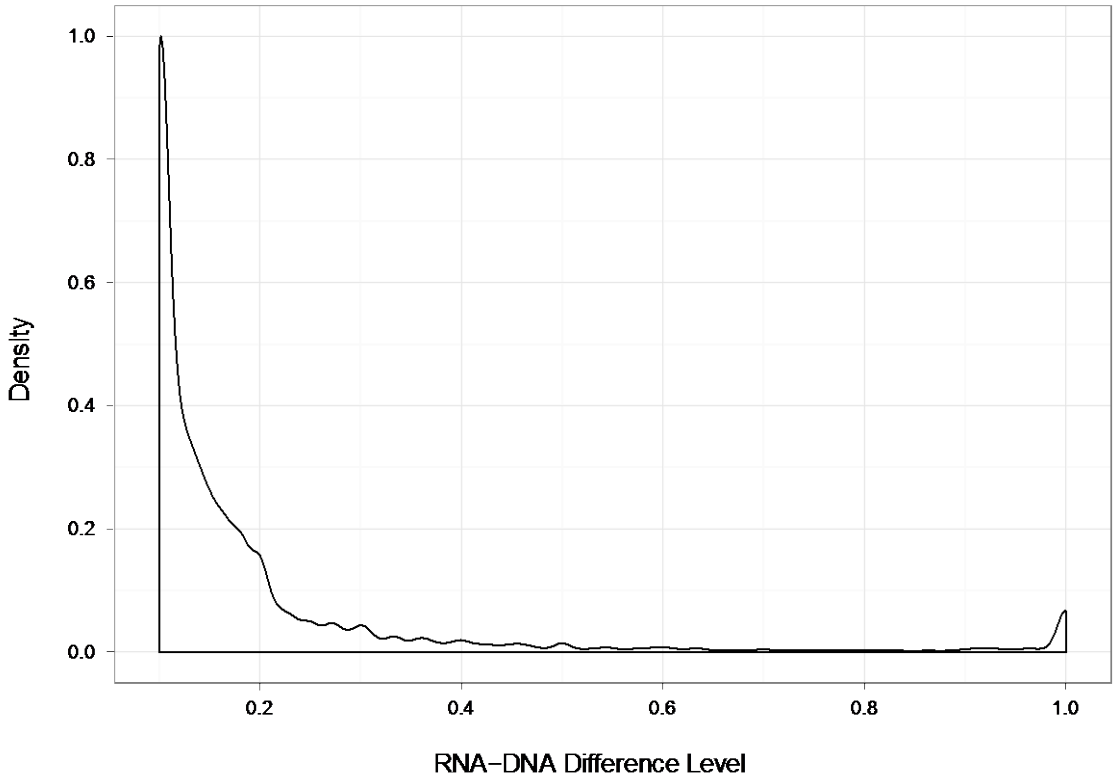
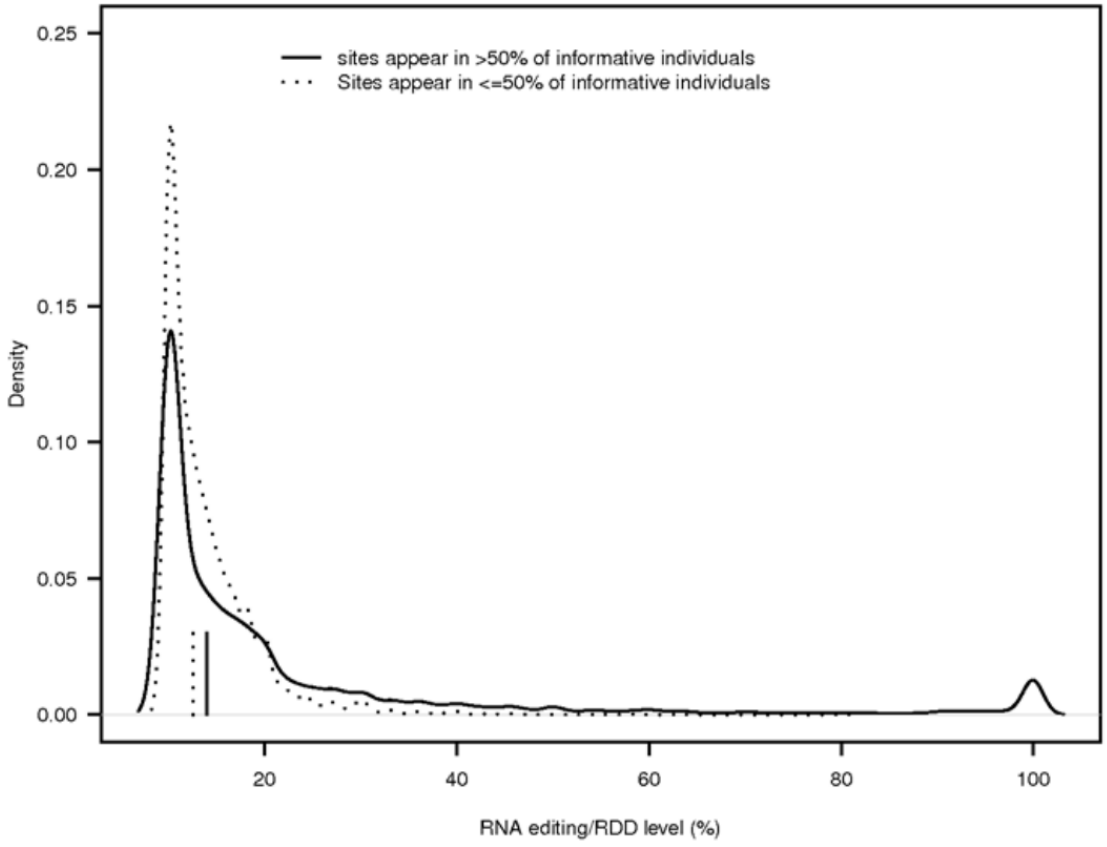


Figure 3.8 Distribution of RDD levels by frequency of event across 27 individuals. The distribution of RDD levels among the 27 unrelated individuals is shown here for sites that appear in the majority versus minority of informative individuals.



3.4 Discussion

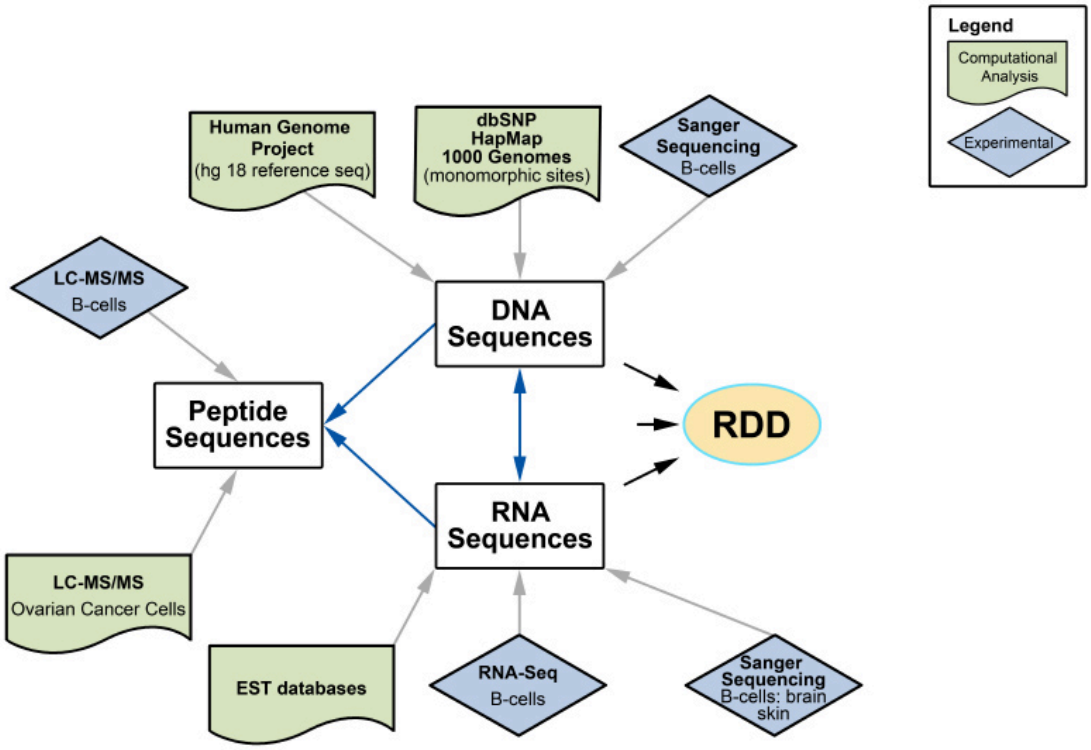
In this study, we have uncovered thousands of exonic sites where the RNA sequences do not match those of the DNA sequences, including transitions and transversions. These findings challenge the long-standing belief that in the same individuals, DNA and RNA sequences are nearly identical. To increase the confidence in our results, we obtained the DNA, RNA, and protein sequences from different individuals and cell types using a range of technologies (Figure 3.9). The samples included cell lines and primary cells from healthy individuals and tumors. We used data from public resources such as EST databases, the HapMap, and 1000 Genomes Project, as well as those that we generated with traditional Sanger sequencing, high-throughput sequencing technologies, and mass spectrometry. We observed sites where the RNA and peptide sequences are the same but differ from the corresponding DNA sequences. The results support our observation that in an individual, DNA and RNA sequences from the same cells are not always identical and some of the variant RNA sequences are translated into proteins. The consistent pattern of the observations suggests that the RDDs have biological significance and are not just “noise”. At nearly all RDD sites, we observed only one RDD type across cell types and in different individuals. If the DNA sequence is A/A, and the RNA is A/C in one sample, in other samples, we see the same A-to-C difference, but not other types of differences. These results suggest that there are unknown aspects of transcription and/or post-transcriptional processing of RNA. These differences may now be studied along with those in other genomes and organisms such as the mitochondrial genomes of trypanosomes and chloroplasts of plants, where RNA

editing and modifications are relatively common (Hundley et al. 2008; Phreaner et al. 1996).

The underlying mechanisms for these events are largely unknown. For most of the cases, we do not know yet whether a different base was incorporated into the RNA during transcription or if these events occur post-transcriptionally. About 23% of the sites are A-to-G differences; some of these are likely mediated by ADAR, but other, currently unknown, mechanisms can be involved. If it is a co-transcriptional process, then the signal can be in the DNA or the RNA such as secondary structures or modified nucleotides. In addition, as some of the RDDs are found near splice and poly(A) sites, it is possible that this may be a facet of systematic RNA processing steps such as splicing and cleavage (Rueter et al. 1995; Rueter et al. 1999).

Our findings supplement previous studies demonstrating sequence differences between DNA and RNA in the human genome and show that these differences go beyond A-to-G transition. These findings affect our understanding of genetic variation; in addition to DNA sequence variation, we identify individual variation in RNA sequences. For monomorphic DNA sequences that show RDD, there is an overall increase in genetic variation. Thus, this variation not only contributes to individual variation in gene expression, but also diversifies the proteome because some identified sites lead to nonsynonymous amino acid changes. We speculate that this RNA sequence variation likely affects disease susceptibility and manifestations. To date, mapping studies have focused on identifying DNA variants as disease susceptibility alleles. Our results suggest that the search may need to include RNA sequence variants that are not in the DNA sequences.

Figure 3.9 Data generated and analyses conducted for RDD study.



3.5 Materials and Methods

3.5.1 Samples

We used data generated by sequencing RNA from 27 cultured B-cell lines obtained from Coriell (Camden, NJ, USA). These cell lines were derived from individuals of European descent from the Utah pedigrees of the CEPH collection, which have been extensively genotyped or sequenced by the International HapMap and the 1000 Genomes Projects. All individuals are unrelated. The list of individuals is shown in Table 3.1. The B-cells were grown to a density of 5 x 10^5 cells/mL in RPMI 1640 supplemented with 15% fetal bovine serum, 100 units/mL penicillin and 100 μg/mL streptomycin, and 2 mmol/L L-glutamine. Foreskin tissues were collected from circumcisions of 10 healthy and unrelated 3-day old newborns (anonymous donors at the Hospital of the University of Pennsylvania). Primary fibroblasts were isolated from these foreskin tissues. Briefly, the tissues were sectioned and the epidermis was removed. The remaining dermis was incubated in 1 ml of 3 mg/ml collagenase (Roche, Indianapolis, IN, USA) in HBSS buffer with calcium and magnesium (Mediatech, Manassas, VA, USA) at 37°C for 30 minutes. 1 ml of 0.5% Trypsin/EDTA (Invitrogen, Carlsbad, CA, USA) was then added, and the tissue was incubated at 37°C for 10 minutes. After inactivating trypsin by adding 1 ml MEM medium (Invitrogen) with 10% fetal bovine serum (Hyclone, Logan, UT, USA), undigested tissue pieces were removed and fibroblasts were collected from suspension by centrifugation. Passage-1 fibroblasts were subsequently cultured at 37°C in 5% CO2 in MEM medium supplemented with 10% fetal bovine serum, 2 mmol/L L-glutamine, 100 units/mL penicillin and 100 μg/ml streptomycin, and were passaged every 3 days. Human brain tissues were obtained from National Disease Research Interchange.

Tissues from cerebral cortex were collected during routine autopsy of cardiac-death donors within 7~12 post-mortem hours. Samples were snap-frozen and kept at -80°C until DNA/RNA extraction. All 10 subjects were Caucasians with age of 43~79 years, including both males and females.

3.5.2 RNA-Sequencing

RNA-Seq was performed as recommended by the manufacturer (Illumina, Hayward, CA, USA). Briefly, cells were harvested 24 hours after addition of fresh medium and total RNA was extracted using the RNeasy Mini kit with DNase treatment (Qiagen, Valencia, CA, USA). PolyA mRNA was isolated and fragmented. First strand cDNA was prepared using random hexamers. Following second strand cDNA synthesis, end repair, addition of a single A base, adaptor ligation, agarose gel isolation of ~200 bp cDNA and PCR amplification of the ~200 bp cDNA, the samples were sequenced using the Illumina 1G Genome Analyzer to a coverage of approximately 41 million 50-bp reads per sample. High quality reads were mapped to the Gencode mRNA sequences (build hg18) using program Bowtie (v0.12.7) (Langmead et al. 2009). We used the default settings of Bowtie (allowing up to two mismatches) and then extracted uniquely mapped reads from the obtained SAM file.

3.5.3 Identification of RNA-DNA sequence differences

We considered RDD events only at sites that are homozygous in the genomic DNA. Such sites can be either a SNP with homozygous genotype or a monomorphic site, which we define as sites that are not identified as SNPs in dbSNP, HapMap or the 1000 Genomes Project. We obtained SNP sites and the corresponding genotypes from the HapMap and the 1000 Genomes Projects. We further required SNPs found in both the

HapMap and 1000 Genomes Project to be concordant. For monomorphic sites, we assumed that all individuals are homozygous for the reference allele in the reference genome. We identified RDD events by comparing genotypes derived from DNA-Seq with those from RNA-Seq on the same individual. For each individual, an RDD event is declared at a site if 1) the number of uniquely mapped reads in RNA-Seq is ≥ 10 and the per-base quality score for each of the bases is ≥ 25, 2) the number of uniquely mapped reads in DNA-Seq is ≥ 4 in the 1000 Genomes Project, 3) the individual's genotype at the site is homozygous, and 4) ≥ 10% of the mapped reads in RNA-Seq show an allele that differs from the individual's (DNA-level) genotype. We removed all sites that overlap with segmental duplications (Bailey et al. 2002; Bailey et al. 2001) and repetitive sequences (defined by RepeatMasker on the UCSC Genome Browser).

3.5.4 EST search for RNA-DNA sequence differences

We analyzed RDD sites where the DNA sequences are monomorphic. For each site, we used BLAST to search for sequences corresponding to the RNA-allele of the RNA-DNA sequence difference and 25 nucleotides up- and downstream of that site in human dbEST. An EST is found to contain the RDD allele if its sequence is identical to the RNA-form of the RDD site and there is no more than 1 mismatch (except for known SNPs) to the 51-nucleotide query sequence. The EST must also map to the gene where the RDD was called (the mapping is mostly according to Unigene).

3.5.5 Sanger sequencing validation of RNA-DNA sequence differences

To validate the sites we identified as RNA-DNA sequence differences by RNA-Sequencing, we carried out Sanger sequencing. Genomic DNA and total RNA were extracted using the Qiagen AllPrep DNA/RNA kit (with DNase treatment for total RNA)

from B-cells and foreskin fibroblast. DNA and RNA from human brain tissues were extracted using Gentra Puregene Tissue Kit and RNeasy Lipid Tissue Mini kit (Qiagen), respectively. RNA was converted to single stranded cDNA using TaqMan reverse transcription reagents with oligo d(T) priming (Applied Biosystems, Foster City, CA, USA). PCR products (see Table 3.9) encompassing putative RDD sites were generated from the cDNA and sequenced using BigDye Terminator Cycle Sequencing on a 3730 DNA Analyzer (Applied Biosystems). RDD sites were validated by visual inspection of sequencing traces independently by at least two individuals.

Table 3.9 Primer sequences used for Sanger sequencing validation of RDDs.

Primer Name	RDD Site	Primer Sequence
DPP7- F	Chr9:139128755	5’- CAGGAGCGCTTCTTCCAGC- 3’
DPP7- R	Chr9:139128755	5’- CCGACACCAGGAAGCGC- 3’
HLA- DQB2- F1	Chr6:32833537	5’- CCTAGGGTGGTCAGACTGGA- 3’
HLA- DQB2- R1	Chr6:32833537	5’- TTGGGTTCCTCACCTTTCTG- 3’
HLA- DQB2- F2	Chr6:32833545	5’- CCTAGGGTGGTCAGACTGGA- 3’
HLA- DQB2- R2	Chr6:32833545	5’- TTGGGTTCCTCACCTTTCTG- 3’
HLA- DQB2- F3	Chr6:32833550	5’- CCTAGGGTGGTCAGACTGGA- 3’
HLA- DQB2- R3	Chr6:32833550	5’- TTGGGTTCCTCACCTTTCTG- 3’
EIF2AK2- F	Chr2:37181512	5'- CCCCAAGAGCCACATGTATT- 3'
	Chr2:37181517	
	Chr2:37181520	
EIF2AK2- R	Chr2:37181512	5'-CCTCAAGCTCACTGTCACCA-3'
	Chr2:37181517	
	Chr2:37181520	
AZIN1- F	Chr8:103910812	5'- TACAAGGAAGATGAGCCTCTG- 3'
AZIN1- R	Chr8:103910812	5'- AATAAATGGCTGGCCTCTGA- 3'
BLCAP- F	Chr20:35580977	5’- CCCGGCAGAGATCATGTATT- 3’
BLCAP- R	Chr20:35580977	5’- AACAGTTTCCCCAGCAGCTA- 3’
PPWD1- F	Chr5:64894960	5’- AGCGATTTTCAGCAGAGACG- 3’
PPWD1- R	Chr5:64894960	5’- TTCCTCTTCTTGGCCAGTGT- 3’
NDUFC2- F	Chr11:77468303	5'-ACCAGGCCTCAAGTGGAAAC- 3’
NDUFC2- R	Chr11:77468303	5'-AGCCGTCGCGATCGG- 3’

3.5.6 Mass spectrometry analysis of proteome

Protein Database with RDD sites

We made a protein database using Gencode mRNA sequences. For genes that display non-synonymous RDDs, protein forms predicted from both DNA sequences and RNA sequences were included. This database consists of 17,726 protein entries, among which 2,057 proteins contain RDD sites. This allows us to search for spectra of peptides encoded by the DNA and RNA forms of mRNAs simultaneously. Mass spectrometry data were searched against this database using a locally installed version of Mascot search engine (Perkins et al. 1999) (Matrix Science, Boston, MA, USA). In order to evaluate the certainty of search results and the false discovery rate, we included a decoy database search. The decoy database contains the reversed amino acid sequences of each protein contained in the original database, and thus no true matches should be expected (Deutsch et al. 2010). Each dataset was searched against both the original and the decoy databases, using the same parameters. The false discovery rate (FDR) was determined using the number of the matches from the original database (TP) and those from decoy database (FDR=FP/(TP+FP)).

Ovarian cancer and leukemic cells

First we mined public databases for large-scale mass spectrometry data on human cells. We found two comprehensive datasets on proteomic profiling of ovarian cancer cells (Sodek et al. 2008) and leukemic cells, K562 (https://proteomecommons.org/dataset.jsp?i=76466). The data were downloaded from Proteome Commons database. The downloaded m/z XML files were converted to Mascot compatible format using ReAdW (Elias et al. 2005). We used the same parameters as the

original study (Sodek et al. 2008) for our Mascot search: Parent ion Δmass of 4 Da, fragment mass error of 0.4 Da, complete carbaminomethyl modification of cysteine, and allow one trypsin miscleavage; for the leukemic cells: parent ion Δmass of 0.3 Da, fragment mass error of 0.5 Da, complete carbaminomethyl modification of cysteine and variable N-terminal acetylation and methionine oxidation, allowing up to one trypsin miscleavage (same as B-cells, see below). False discovery rate is < 1% (determined by Mascot decoy database search).

B-cells

Cultured B-cells from 15 CEPH indivduals were cultured as described above (see section 3.4.1) for GEL/LC-MS/MS. Equal number of cells from each individual were pooled and lysed in 20 mM Tris HCl pH 8, 137 mM NaCl, 10% glycerol, 1% Nonidet P-40 (NP-40) and 2 mM EDTA. 60 µg of total protein was pre-fractionated and digested as described by Beer and colleagues (Beer et al. 2011). Specifically, the sample of 60 µg of total protein was loaded onto a 12% NuPAGE Bis-Tris gel and was run until the dye reached 2 cm from the bottom of the loading well. The gel was stained with Coomassie R250 and de-stained using methanol and water. The entire lane of protein was sliced into 27 1-mm slices using a razor array. Gel slices were destained in 200 µl of 200 mM ammonium bicarbonate and 50% acetonitrile for 30 minutes with shaking at 37°C and then dried in a Speedvac. This was followed by incubation in 100 µl of 20 mM TCEP in 25 mM ammonium bicarbonate (pH 8.0) for 15 minutes at 37°C. The supernatant was discarded and 100 µl of 40 mM iodoacetamide in 25 mM ammonium bicarbonate (pH 8.0) was added and incubated for 30 minutes at 37°C. The supernatant was again discarded and the gel slices were washed twice with 200 µl of 25 mM ammonium

bicarbonate in H2O and once with 25 mM ammonium bicarbonate in 50% acetonitrile. The gel slices were dried and then rehydrated with 20µl of 0.02 µg/µl modified Trypsin (Promega, Madison, WI, USA) in 40 mM ammonium bicarbonate overnight with shaking at 37°C. The supernatant was then transferred to a clean tube and the gel slices were incubated in 20 µl of 40 mM ammonium bicarbonate for 30 min at 37°C. The supernatants were combined and 4 µl of acetic acid was added. Extracted tryptic peptides were injected onto a nanocapillary reverse-phase column (75 µm column terminating in a nanospray 15 µm tip) directly coupled to a LTQ-Orbitrap mass spectrometer (Thermo Scientific, Pittsburgh, PA, USA). The MS/MS data was acquired using a top six method. Each fraction of tryptic peptides was injected and analyzed in triplicates to increase the depth of protein detection. The raw spectra files were converted to Mascot compatible format using BioWorks v 3.3.1 (Thermo Scientific, Waltham, MA, USA). We used the same protein database as in the cancer cell analysis described above. The following Mascot search parameters were applied: parent ion Δmass of 0.3 Da, fragment mass error of 0.5 Da, complete carbaminomethyl modification of cystein and variable N-terminal acetylation and methionine oxidation, allowing up to one trypsin miscleavage. False discovery rate is <1% (determined by Mascot decoy database search); see Table S9 for details on search results. At FDR < 1%, 38,572 unique peptides (from 3,217 proteins) were detected. Among them, 11,496 peptides are from the 659 proteins that contain one or more RDD sites although the RDD-containing peptides were not necessarily detected. Among the 659 proteins, we detected 299 and 28 peptides encoded by the DNA and RNA forms of RDDs. As shown in Table 3.6, 17 pairs of peptides differ in one residue that corresponded to the RNA and DNA variants at RDD sites. An RDD site in the RPL28

protein led to a loss of a STOP codon and resulted in a larger protein with additional amino acids. We carried out BLAST search and ensured that all 28 peptides that correspond to the RNA forms of the RDD-containing peptides are unique to the proteins of interests. For these alignments, we used “nr” to search all non-redundant sequences (which includes CDS translations + PDB + SwissProt + PIR + PRF). All 28 RDD sites correspond to monomorphic sites; by Sanger sequencing of the same aliquot of B-cells used for mass spectrometry, we validated the sequences correspond to the reference genome sequence (Table 3.10).

Table 3.10 Primer sequences used for validating the DNA sequences of RDD sites found in peptides.

Primer Name	RDD Site	Primer Sequence
CD22_F	Chr19:40514815	5'- CCCTGCTCAGGCTTGCACCC- 3'
CD22_R	Chr19:40514815	5'- GCTGCCCCCACCCTACCCTA- 3'
DFNA5_F	Chr7:24705225	5'- AGGTTTGGGATTGTGCAGCGCT- 3'
DFNA5_R	Chr7:24705225	5'- AGCCTTGGCCAGTAACACGTACT- 3'
ENO1_F	Chr1:8848125	5’- ACAGTCCCTGTGTAGCAGCTGT- 3’
ENO1_R	Chr1:8848125	5’- GTCAGCCAGCTGGTCAGGCG- 3’
FH_F	Chr1:239747217	5'- AGGCAAGCCAAAATTCCTTCCGGA- 3'
FH_R	Chr1:239747217	5'- GCCAACATTTCCACAAATGCCACT- 3'
HMGB1_F1	Chr13:29935772	5'- GGAGATCCTAAGAAGCCGAGAGGCA- 3'
HMGB1_R1	Chr13:29935772	5'- AGCGTCCCACTACGAGAATGCCA- 3'
HMGB1_F2	Chr13:29935469	5'- TGGCATTCTCGTAGTGGGACGCT- 3'
HMGB1_R2	Chr13:29935469	5'- ACGGAGGCCTCTTGGGTGCA- 3'
ITPR3_F	Chr6:33755773	5'- GTGTGACACGTGCCCCCTCC- 3'
ITPR3_R	Chr6:33755773	5'- TGTCCCCTGGCCTCCGGTTC- 3'
RAD50_F	Chr5:131979610	5'- GCACTTGCTGTCACCAGTTGCC- 3'
RAD50_R	Chr5:131979610	5'- TCTGTGCAGCAGGCTAGCAGA- 3'
ROD1_F	Chr9:114026264	5'- TGCGGCCCAGGTAGTTGACTC- 3'
ROD1_R	Chr9:114026264	5'- AGCGCCATTTTGCGATCTTTCCTG- 3'
RPL32_F	Chr3:12852658	5'- TGGTGCTGGGACTCATTGCCT- 3'
RPL32_R	Chr3:12852658	5'- TCTTCACTGCGCAGCCTGGC- 3'
RPS25P8_F	Chr11:118393375	5'- AGGGAAAGGGTGCTTCTGCCA- 3'
RPS25P8_R	Chr11:118393375	5'- GAGCTCCTGAAGGGCTGCCC- 3'
RPS3AP47_F	Chr4:152243651	5'- TGCTTCGTCTGTTCTGTGTTGGTT- 3'
RPS3AP47_R	Chr4:152243651	5'- CAAGGTTGTGTTGTGTGAGGAAGCA- 3'
SUPT5H_F	Chr19:44655806	5'- TGGCTCCCAGACGCCCATGT- 3'
SUPT5H_R	Chr19:44655806	5'- TCACCGTGACGGCGTGTTGG- 3'
TOR1AIP1_F	Chr1:178144365	5'- ACCTGCTTTGCTGTAGGAAGTGGT- 3'
TOR1AIP1_R	Chr1:178144365	5'- TGGGGCCCATTCCTGGGGAG- 3'
TUBA1_F	Chr2:219823379	5'- GCTGAGCAACACGACCGCCA- 3'
TUBA1_R	Chr2:219823379	5'- GCAGCAGCAGCATGAAGGGGA- 3'

Chapter 4. Detection Theory in Identification of RNA-DNA Sequence Differences Using RNA-Sequencing

4.1 Abstract

Advances in sequencing technology have allowed for detailed analyses of the transcriptome at single-nucleotide resolution, facilitating the study of RNA editing or sequence differences between RNA and DNA genome-wide. In humans, two types of post-transcriptional RNA editing processes are known to occur: A-to-I deamination by ADAR and C-to-U deamination by APOBEC1. In addition to these known sequence differences, researchers have reported the existence of all 12 types of RNA-DNA sequence differences (RDDs). However, the validity of these claims is debated, as many studies assert that technical artifacts account for the majority of these noncanonical sequence differences. In this chapter, we use a detection theory approach to evaluate the performance of RNA-Sequencing (RNA-Seq) and associated aligners in accurately identifying RDDs. By generating simulated RNA-Seq datasets with RDDs inserted in known locations, we assessed the sensitivity and false discovery rate of RDD detection. Overall, we found that alignment errors do not significantly influence RDD discovery in the absence of sequencing error, as the false negative and false discovery rates of RDD detection can be contained below 10% with minimal thresholds. We evaluated the impact of sequencing error on the false discovery rate and found that the effect of random sequencing errors can be mitigated with stricter thresholds on RDD identification. In contrast, non-random sequencing errors that occur at high levels cannot be distinguished from true positive RDDs and play a nontrivial impact on the false discovery rate of RDD detection. We also determined the performance of various filters that target false positive

RDDs and found them to be effective in discriminating between true and false positives. Lastly, we used the optimal thresholds and parameters we identified from our synthetic analyses to identify RDDs in a human lymphoblastoid cell line. We found approximately 9,000 RDDs, the majority of which are A-to-G edits and likely to be mediated by ADAR. Moreover, we found the majority of non A-to-G RDDs to be associated with poorer alignments and determine that the evidence for noncanonical RDDs in humans to be weak. Overall, we found RNA-Seq to be a powerful technique when coupled with the appropriate thresholds and filters for surveying RDDs genome-wide. We aim for this work to provide a practical framework for those interested in the study of site-specific allelic differences genome-wide using high-throughput sequencing.

4.2 Introduction

Next-generation sequencing technology provides comprehensive sequence information. The precision afforded by RNA-Sequencing (RNA-Seq) is useful for studying various aspects of the transcriptome such as alternative splicing (Pan et al. 2008; E. T. Wang et al. 2008), RNA editing (Park et al. 2012; Peng et al. 2012), and differential allelic expression (DeVeale et al. 2012; Gregg et al. 2010; Heap et al. 2010; R. D. Lee et al. 2013). RNA editing refers to co- or post-transcriptional modification of RNA, resulting in a transcript that is different from the underlying genomic template. In humans, two types of RNA editing processes are known to occur: adenosine deamination by ADAR results in A-to-G edits (Bass & Weintraub 1988; Nishikura 2010) and cytidine deamination by APOBEC1 results in C-to-U changes (S. H. Chen et al. 1987; Smith et al. 2012).

In recent years, many genome-wide surveys of RNA editing in humans have been performed using next-generation sequencing technology (Bahn et al. 2012; Kleinman et al. 2012; Levanon et al. 2004; J. B. Li et al. 2009; Peng et al. 2012). In addition to the known A-to-G and C-to-U alterations introduced by RNA editing, researchers have reported the existence of RNA-DNA sequence differences (RDDs) that cannot be explained by known mechanisms (Bahn et al. 2012; Ju et al. 2011; M. Li et al. 2011). However, the validity of these results is contested, with many reports citing experimental and technical artifacts as the main determinants of systematic sequence differences between RNA and DNA (Kleinman et al. 2012; Kleinman & Majewski 2012; W. Lin et al. 2012; Pickrell et al. 2012; Schrider et al. 2011).

Previous approaches for accurate identification of RDDs mainly involved ad hoc filters aimed at removing false positives (Bahn et al. 2012; Kleinman et al. 2012; Ramaswami et al. 2012). In this study, we use a detection theory approach to evaluate the relative effect of misalignment and sequencing error on RDD analysis. In particular, we generated synthetic RNA-Seq datasets containing simulated RDDs at known locations and assessed the performance of various RNA-Seq aligners in accurately identifying RDDs. We also analyzed filtering methods for their efficacy in minimizing the false discovery rate of RDD detection while simultaneously maintaining high sensitivity values. To complement our *in silico* analyses, we assessed the sequencing error profile of a dataset comprising human cDNA clones deeply sequenced with Illumina Hi-Seq technology and evaluated the effect of non-random errors on the false discovery rate of RDD detection. Lastly, after determining the optimal thresholds and parameters for sequence difference analysis, we searched for the presence of RDDs in an experimental human RNA-Seq dataset for which deep DNA and RNA sequence information is publicly available.

Overall, our report aims to explore the phenomenon of RDDs in humans as well as provide a framework for those interested in the study of RNA editing, RDDs, or differential allelic expression by elucidating the appropriate thresholds and parameters for accurate detection of allele-specific differences in RNA-Seq data. The synthetic datasets generated in this study are available at the GEO public repository.

4.3 Results

4.3.1 Simulated RNA-Seq datasets

To evaluate the performance of various alignment algorithms and filtering methods in detecting RDDs, we generated synthetic RNA-Seq datasets containing simulated RDDs at known locations (see Materials and Methods). First, we created a "clean" dataset (dataset 1) with no sequencing errors or intronic reads in order to evaluate the degree of bias introduced by alignment error alone. Next, in order to capture the effect of sequencing error on RDD identification, we generated a more realistic RNA-Seq dataset containing substitutional sequencing errors, indel polymorphisms, intronic signal, and lower quality bases at the tail end of reads. For our initial analyses, we assume a simplistic sequencing error model in which misincorporations occur randomly and independently at the same rate. Later, we consider the effect of sequencing error profiles derived from experimental Illumina Hi-Seq datasets on RDD detection. Both datasets contain 50 million pairs of non strand-specific reads of length 100 base pairs (bp) and were generated in triplicates to allow for assessment of variability of our various metrics.

Datasets were aligned using GSNAP, RUM, and Tophat2 (see Materials and Methods). For both datasets, GSNAP performed the best in terms of the number of reads mapped in total and uniquely (Table 4.1), aligning approximately 99% of the 50 million read pairs. In contrast, RUM and Tophat2 aligned approximately 98% of the read pairs in dataset 1, but only roughly 95% in dataset 2, which contains sequencing errors. Overall, between 97 to 99% of the read pairs are aligned uniquely with GSNAP and RUM, whereas only approximately 89% of the read pairs are aligned uniquely with Tophat2.

Table 4.1 Alignment statistics of simulated RNA-Seq datasets.

Dataset	Aligner	Statistic	Value
dataset 1	gsnap	Number of Read Pairs Aligned	49706613 ± 5861
		Percentage of Read Pairs Aligned	99.41 ± 0.012%
		Number of Read Pairs Aligned Uniquely	49057838 ± 5548
		Percentage of Read Pairs Aligned Uniquely	98.69 ± 0.0012%
dataset 1	rum	Number of Read Pairs Aligned	49338154 ± 5464
		Percentage of Read Pairs Aligned	98.68 ± 0.011%
		Number of Read Pairs Aligned Uniquely	48192283 ± 5415
		Percentage of Read Pairs Aligned Uniquely	97.68 ± 0.0039%
dataset 1	tophat2	Number of Read Pairs Aligned	49144986 ± 6939
		Percentage of Read Pairs Aligned	98.3 ± 0.013%
		Number of Read Pairs Aligned Uniquely	43744298 ± 5756
		Percentage of Read Pairs Aligned Uniquely	89.01 ± 0.0084%
dataset 2	gsnap	Number of Read Pairs Aligned	49790494 ± 3376
		Percentage of Read Pairs Aligned	99.58 ± 0.0068 %
		Number of Read Pairs Aligned Uniquely	49095402 ± 4246
		Percentage of Read Pairs Aligned Uniquely	98.6 ± 0.0034 %
dataset 2	rum	Number of Read Pairs Aligned	47975774 ± 34932
		Percentage of Read Pairs Aligned	95.95 ± 0.07 %
		Number of Read Pairs Aligned Uniquely	46820152 ± 35128
		Percentage of Read Pairs Aligned Uniquely	97.59 ± 0.0062 %
dataset 2	tophat2	Number of Read Pairs Aligned	47763421 ± 77673
		Percentage of Read Pairs Aligned	95.61 ± 0.16 %
		Number of Read Pairs Aligned Uniquely	43072870 ± 73773
		Percentage of Read Pairs Aligned Uniquely	90.18 ± 0.015 %

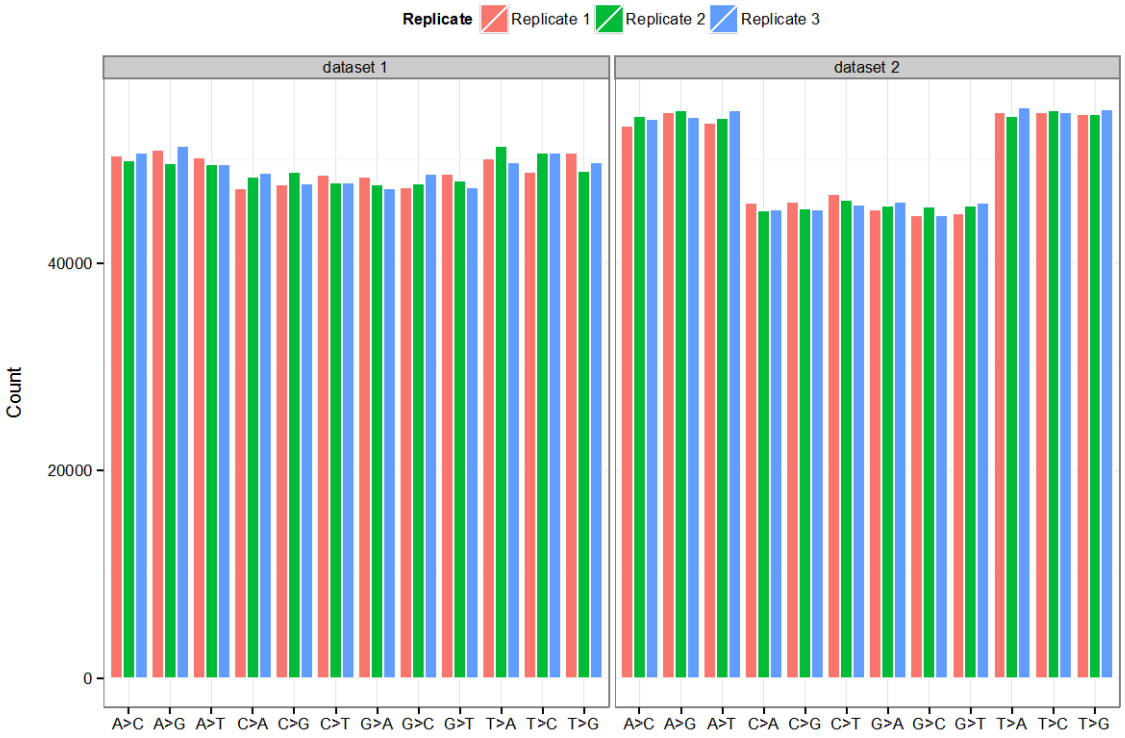
4.3.2 Simulated RNA-DNA sequence differences

For each of the two datasets, we inserted RDDs at random known locations in the genome (see Materials and Methods). In particular, at positions containing RDDs, a fraction of the total reads at the site bear a randomly chosen non-reference allele representing the sequence difference. Furthermore, we define the fraction of reads containing the non-reference base to be the RDD level.

We generated approximately 600,000 total RDDs on average for both datasets. Each RDD type is equally represented, with sequence differences that originate from cytosine and guanosine (C>A, C>G, C>T, G>A, G>C, G>T) slightly overrepresented than other types. This variation results from differing base compositions throughout the genome, with the effect more pronounced in dataset 2, which contains reads from intronic regions of the genome (Figure 4.1).

The coverage, or the total number of reads, at a given site is important in the analysis of RDDs as the presence and levels of RDDs at sites that are deeply sequenced are more likely to be robustly assessed. In order to assess the effect of sequencing depth on RDD detection, we stratified sites in the genome according to coverage and simulated equal numbers of RDDs in each group (see Materials and Methods).

Figure 4.1 Total number of simulated RNA-DNA sequence differences. For both datasets, approximately 600,000 RNA-DNA sequence differences were generated in each replicate. Differences in the number of each type of RNA-DNA sequence difference reflect underlying variation in base composition throughout the genome, as dataset 2 contains reads originating from intronic regions whereas dataset 1 does not. Colors refer to the three replicates generated for each dataset.



For each RDD, we chose the level, or proportion of reads carrying the non-reference base, from a standard uniform random distribution excluding 0. However, because of the discrete nature of coverage, the distribution of RDD levels is not uniform at sites with low coverage; for sites with coverage greater than 100, the distribution of levels is uniform across all levels with the exception of boundary values (Figure 4.2).

To understand the effect of hyperediting by ADAR and the observation that non-canonical RDDs often cluster (Bahn et al. 2012), we modeled a subset of RDDs to occur in close proximity of one another (see Materials and Methods). Hyperediting refers to a type of editing by members of the ADAR family whereby approximately 50% of the adenosines on each strand of an RNA duplex is edited in a promiscuous fashion (Polson & Bass 1994). For each dataset, we generated approximately 2,000 clusters of length 100 bp within which approximately 50% of all positions bear the same type of RDD (see Materials and Methods).

Overall, in dataset 1, the average distance between neighboring RDDs is 10 bp (median 3 bp) for sites belonging to clusters and 815 bp (median of 58 bp) for those that do not. For dataset 2, which contains intronic reads, the average distance between RDDs is also 10 bp (median 3 bp) for sites belonging to clusters and 1565 bp (median 225 bp) for those that do not (Table 4.2).

Figure 4.2 Levels of simulated RNA-DNA sequence differences. Here we depict the distribution of RDD difference levels, or the proportion of reads at the sequence difference site that carry the sequence difference. Because of the discrete nature of RNA-Seq data, the levels of RDDs at sites with relatively low coverage is not uniform as shown by the blue area, which represents sites with coverage less than 10. For sites with coverage greater than 100 (red area), the density curve of sequence difference levels is fairly uniform except at boundary conditions.

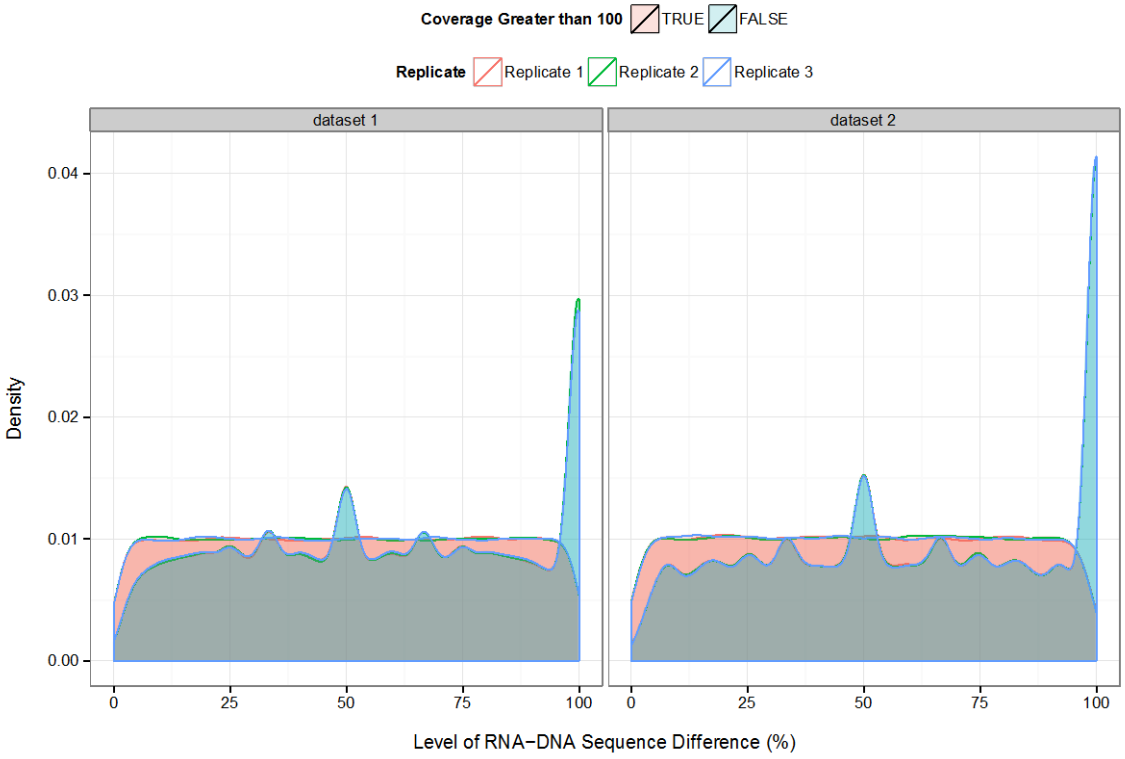


Table 4.2 Summary statistics on distance between neighboring RNA-DNA sequence differences.

Dataset	In Cluster	Number of Sites	Distance to nearest RNA-DNA sequence difference*					
			Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
dataset 1	TRUE	219239	1	1	3	10	5	155900
dataset 1	FALSE	369416	1	23	58	815	146	1322700
dataset 1	TOTAL	588654	1	4	19	515	82	1322700
dataset 2	TRUE	222509	1	1	3	10	5	112347
dataset 2	FALSE	376504	1	75	225	1565	869	2317667
dataset 2	TOTAL	599013	1	4	59	987	353	2317667

*Note: Each statistic is averaged across the 3 replicates.

4.3.3 Sensitivity of RNA-DNA sequence difference detection

We began our assessment of the performance of next-generation sequencing technology in identifying RDDs by analyzing the sensitivity or recall rate of sequence difference identification. Here, we focus on the false negative rate of RDD detection and in subsequent analyses, we evaluate the issue of false positives. We begin by defining a simulated RDD as being properly identified by the aligner if at least one read bearing the non-reference base is mapped; later we analyze the correlation between the simulated and observed RDD levels. For dataset 1, we observed that overall, GSNAP detected 96.32 ± 0.06% of the simulated RDDs, whereas RUM and Tophat2 correctly identified 95.36 ± 0.16% and 95.04 ± 0.14%, respectively. For dataset 2, which contains sequencing errors and intronic reads, GSNAP identified 93.54 ± 0.10% of all simulated sites, whereas RUM and Tophat2 found 91.12 ± 0.13% and 90.84 ± 0.15%, respectively.

Next, we investigated the effect of sequencing depth or coverage on the detection of RDDs. We observed that for both datasets, the sensitivity of detection increases with higher thresholds on the minimum depth of coverage (Figure 4.3). For example, the sensitivity of sequence difference detection using GSNAP increases from 96.32 ± 0.06% to 98.73 ± 0.01% in dataset 1 and even more sharply from 93.54 ± 0.10% to 98.03 ± 0.04% in dataset 2 when sites with coverage lower than 10x are removed from consideration. In particular, the sensitivity of detection for sites with coverage less than 10x is 86.11 ± 0.09% in dataset 1 and 80.38 ± 0.30% in dataset 2, versus 97.05 ± 0.14% in dataset 1 and 97.01 ± 0.03% for those with coverage between 10x and 20x. The sensitivity of RDD detection using RUM and Tophat2 increases in a similar fashion with higher coverage (Table 4.3). Given the relatively low recall rate or high false negative

rate of RDD detection for locations with low coverage, we restrict subsequent analyses to sites with a minimum of 10 reads in the simulated dataset and the corresponding aligned datasets per GSNAP, RUM, or Tophat2.

Figure 4.3 Sensitivity of RNA-DNA sequence difference detection versus coverage. The sensitivity or recall rate of RDD identification is shown versus various thresholds on the minimum depth of coverage required at the site of the simulated difference. For all three aligners, the true positive rate increases sharply upon raising the minimum depth of coverage required for detection from 0 to approximately 50, after which the recall rate levels off. Colors refer to the aligner used for RDD detection.

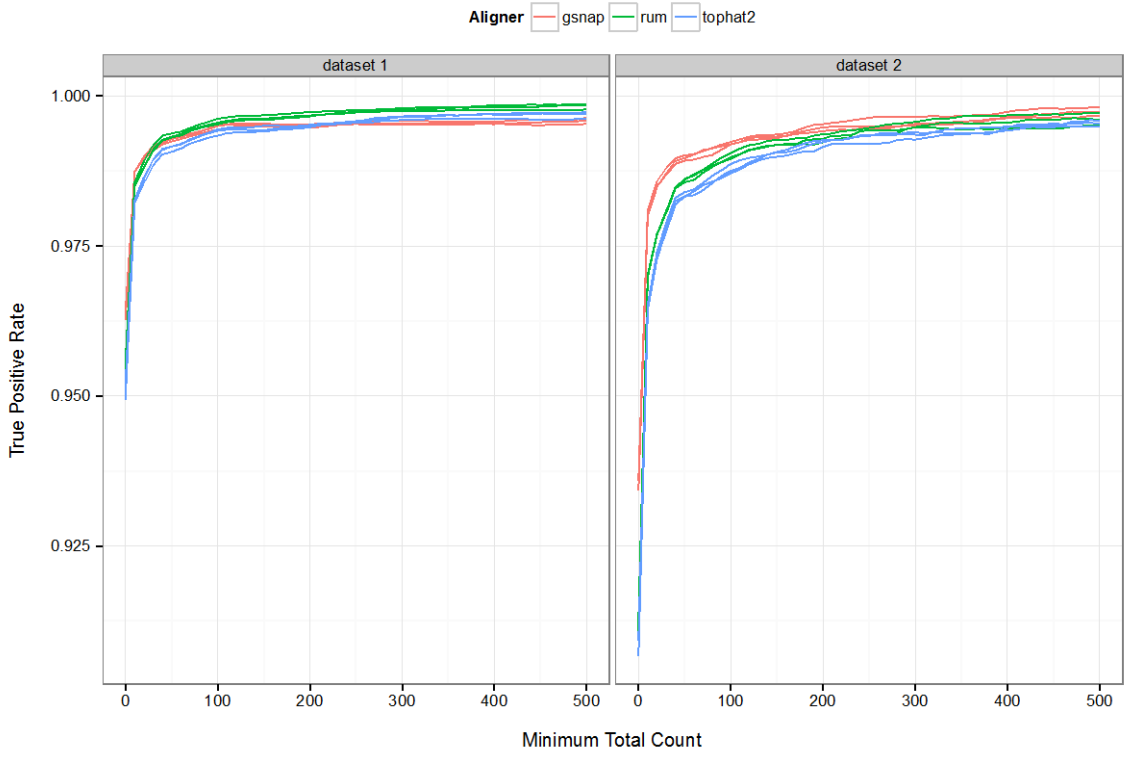


Table 4.3 Sensitivity of RNA-DNA sequence difference detection versus coverage.

Dataset	Aligner	Minimum Coverage or Total Count§	Sensitivity
dataset 1	gsnap	0	96.32 ± 0.06%
dataset 1	rum	0	95.36 ± 0.16%
dataset 1	tophat2	0	95.04 ± 0.14%
dataset 1	gsnap	10	98.73 ± 0.01%
dataset 1	rum	10	98.52 ± 0.03%
dataset 1	tophat2	10	98.24 ± 0.03%
dataset 1	gsnap	50	99.27 ± 0.03%
dataset 1	rum	50	99.34 ± 0.04%
dataset 1	tophat2	50	99.13 ± 0.05%
dataset 1	gsnap	100	99.49 ± 0.04%
dataset 1	rum	100	99.57 ± 0.05%
dataset 1	tophat2	100	99.40 ± 0.06%
dataset 2	gsnap	0	93.54 ± 0.10%
dataset 2	rum	0	91.12 ± 0.13%
dataset 2	tophat2	0	90.84 ± 0.15%
dataset 2	gsnap	10	98.03 ± 0.04%
dataset 2	rum	10	96.99 ± 0.03%
dataset 2	tophat2	10	96.49 ± 0.05%
dataset 2	gsnap	50	98.97 ± 0.04%
dataset 2	rum	50	98.60 ± 0.03%
dataset 2	tophat2	50	98.35 ± 0.04%
dataset 2	gsnap	100	99.21 ± 0.02%
dataset 2	rum	100	98.99 ± 0.05%
dataset 2	tophat2	100	98.78 ± 0.08%

* Note: An RDD is considered properly identified if a minimum of 1 read bearing the sequence difference is present per the aligner.

§ The threshold on the minimum coverage applies to the true coverage at the site of the underlying RDD per the simulated RNA-Seq dataset, not the observed coverage per the aligner.

Next, we analyzed the effect of RDD level on the sensitivity of RDD detection. We binned the simulated sequence differences into 10 groups by RDD levels and evaluated the true positive rate for each group. We found that RDDs with levels between 0 and 10% had the lowest recall rates, ranging from 93.05 ± 0.15% for Tophat2 to 94.87 ± 0.05% for GSNAP in dataset 1 and 86.67 ± 0.14% for Tophat2 to 91.03 ± 0.06% for GSNAP in dataset 2 (Table 4.4). However, for RDDs with a minimum level of 10%, the recall rates are fairly high, averaging 99.11 ± 0.01% in dataset 1 and 98.67 ± 0.04% in dataset 2 for GSNAP. Among the 3 aligners, GSNAP had the highest sensitivity values across all levels (Figure 4.4). Given the lower recall rates for sequence differences with low levels, we restrict our downstream analyses to sites with a minimum level of 10%.

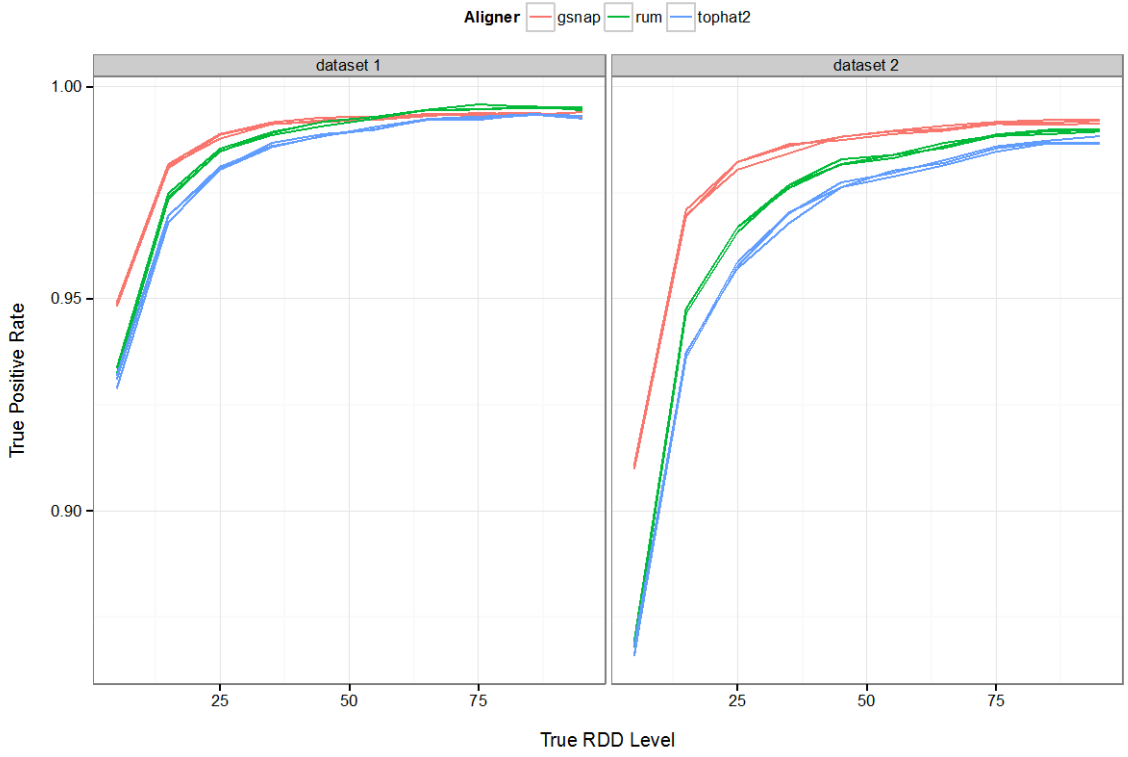
Table 4.4 Sensitivity of RDD detection versus the level of sequence difference.

Dataset	Aligner	RDD Level (%)	Sensitivity
dataset 1	gsnap	0 - 10	94.87 ± 0.05%
dataset 1	rum	0 - 10	93.26 ± 0.08%
dataset 1	tophat2	0 - 10	93.05 ± 0.15%
dataset 1	gsnap	10 - 20	98.12 ± 0.04%
dataset 1	rum	10 - 20	97.41 ± 0.07%
dataset 1	tophat2	10 - 20	96.91 ± 0.09%
dataset 1	gsnap	20 - 30	98.84 ± 0.06%
dataset 1	rum	20 - 30	98.50 ± 0.03%
dataset 1	tophat2	20 - 30	98.07 ± 0.04%
dataset 1	gsnap	30 - 40	99.13 ± 0.02%
dataset 1	rum	30 - 40	98.90 ± 0.04%
dataset 1	tophat2	30 - 40	98.61 ± 0.05%
dataset 1	gsnap	40 - 50	99.21 ± 0.06%
dataset 1	rum	40 - 50	99.14 ± 0.06%
dataset 1	tophat2	40 - 50	98.85 ± 0.03%
dataset 1	gsnap	50 - 60	99.24 ± 0.03%
dataset 1	rum	50 - 60	99.27 ± 0.02%
dataset 1	tophat2	50 - 60	99.01 ± 0.04%
dataset 1	gsnap	60 - 70	99.34 ± 0.03%
dataset 1	rum	60 - 70	99.45 ± 0.01%
dataset 1	tophat2	60 - 70	99.23 ± 0.01%
dataset 1	gsnap	70 - 80	99.35 ± 0.04%
dataset 1	rum	70 - 80	99.50 ± 0.07%
dataset 1	tophat2	70 - 80	99.25 ± 0.04%
dataset 1	gsnap	80 - 90	99.36 ± 0.03%
dataset 1	rum	80 - 90	99.52 ± 0.02%
dataset 1	tophat2	80 - 90	99.34 ± 0.01%
dataset 1	gsnap	90 - 100	99.32 ± 0.07%
dataset 1	rum	90 - 100	99.48 ± 0.04%
dataset 1	tophat2	90 - 100	99.29 ± 0.04%
dataset 2	gsnap	0 - 10	91.03 ± 0.06%
dataset 2	rum	0 - 10	86.87 ± 0.08%
dataset 2	tophat2	0 - 10	86.67 ± 0.14%
dataset 2	gsnap	10 - 20	97.00 ± 0.08%
dataset 2	rum	10 - 20	94.72 ± 0.07%
dataset 2	tophat2	10 - 20	93.68 ± 0.07%
dataset 2	gsnap	20 - 30	98.16 ± 0.10%
dataset 2	rum	20 - 30	96.65 ± 0.07%
dataset 2	tophat2	20 - 30	95.79 ± 0.08%
dataset 2	gsnap	30 - 40	98.56 ± 0.11%
dataset 2	rum	30 - 40	97.66 ± 0.04%
dataset 2	tophat2	30 - 40	96.94 ± 0.14%

Dataset	Aligner	RDD Level (%)	Sensitivity
dataset 2	gsnap	40 - 50	98.79 ± 0.05%
dataset 2	rum	40 - 50	98.21 ± 0.07%
dataset 2	tophat2	40 - 50	97.66 ± 0.07%
dataset 2	gsnap	50 - 60	98.93 ± 0.04%
dataset 2	rum	50 - 60	98.36 ± 0.04%
dataset 2	tophat2	50 - 60	97.95 ± 0.07%
dataset 2	gsnap	60 - 70	99.01 ± 0.06%
dataset 2	rum	60 - 70	98.61 ± 0.06%
dataset 2	tophat2	60 - 70	98.21 ± 0.06%
dataset 2	gsnap	70 - 80	99.14 ± 0.02%
dataset 2	rum	70 - 80	98.85 ± 0.02%
dataset 2	tophat2	70 - 80	98.53 ± 0.06%
dataset 2	gsnap	80 - 90	99.16 ± 0.05%
dataset 2	rum	80 - 90	98.93 ± 0.05%
dataset 2	tophat2	80 - 90	98.69 ± 0.03%
dataset 2	gsnap	90 - 100	99.18 ± 0.05%
dataset 2	rum	90 - 100	98.97 ± 0.03%
dataset 2	tophat2	90 - 100	98.72 ± 0.10%

* Note: An RDD is considered properly identified if a minimum of 1 read bearing the sequence difference is present per the aligner. No threshold on the coverage is imposed.
* Note: Sites with a true coverage value less than 10 per the simulated RNA-Seq dataset are removed from consideration.

Figure 4.4 Sensitivity of RDD detection versus the sequence difference level. Here we depict the true positive rate of RDD detection versus the level of sequence difference, or the proportion of reads at the site bearing the sequence difference allele. A minimum of 1 read bearing the RDD allele is sufficient for a site to be deemed correctly identified. Sites with fewer than 10 reads per the simulated RNA-Seq dataset are removed from consideration. Colors refer to the aligner used for RDD detection.



Next, we analyzed the effect of the repetitive nature of the sequence flanking the sequence difference site on the detection of RDDs. First, we evaluated the sensitivity of detection in regions of the genome that are deemed non-unique by BLAT (see Materials and Methods). We observed that the sensitivity of sequence difference detection using GSNAP in non-unique regions according to BLAT is lower than that in unique regions by approximately 5%. The average sensitivity in non-unique versus unique regions across the three replicates is 94.98 ± 0.11% versus 99.53 ± 2.8 x 10^{-3}% for dataset 1 and 94.74 ± 0.14% versus 99.25 ± 2.6 x 10^{-2}% for dataset 2. The difference in sensitivity of detection between RDDs within non-unique versus unique regions using Tophat2 is similar to GSNAP, while interestingly that for RUM is less than approximately 1% (Figure 4.5 and Table 4.5). We also examined the sensitivity of RDD identification for sites lying within versus outside of RepeatMasker regions (Smit 1996) and observed that across the three replicates, the sensitivity of detection for any of the three aligners is approximately 1 to 2% higher for those lying outside of RepeatMasker regions (Table 4.6).

Lastly, we analyzed the effect of proximity to neighboring RDDs on sensitivity of detection. Short-read aligners typically have a limit on the number of mismatches relative to the reference permitted in a reported alignment, and thus sites with many neighboring sequence differences may be harder to identify. We observed that the sensitivity of sequence difference detection for sites that are greater than 10 bp in distance away from a neighboring sequence difference is roughly 1 to 3% higher for dataset 1 and 3 to 6% higher for dataset 2 (Table 4.7).

Figure 4.5 Sensitivity of RDD detection versus uniqueness of flanking genomic sequence by BLAT.

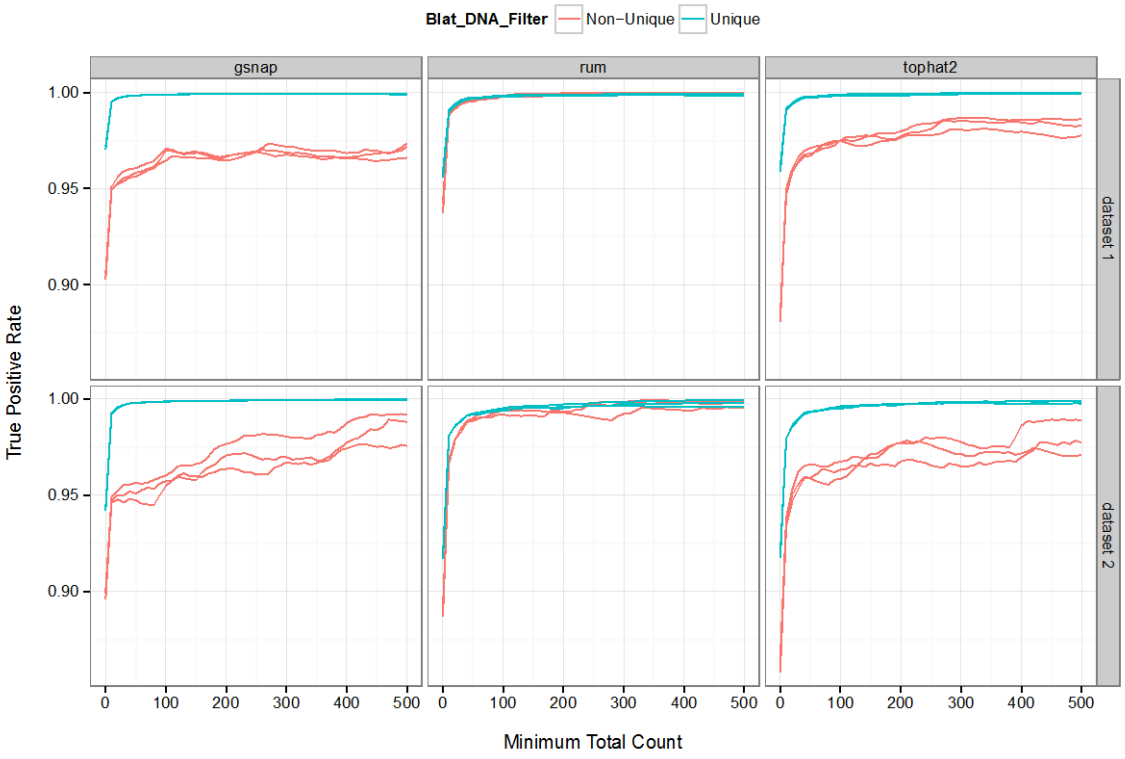


Table 4.5 Sensitivity of RDD detection in unique versus non-unique regions as determined by BLAT.

Dataset	Aligner	Uniqueness	Sensitivity
dataset 1	gsnap	Non-Unique	94.98 ± 0.11%
dataset 1	gsnap	Unique	99.53 ± 0.00%
dataset 1	rum	Non-Unique	98.82 ± 0.07%
dataset 1	rum	Unique	99.05 ± 0.04%
dataset 1	tophat	Non-Unique	94.90 ± 0.18%
dataset 1	tophat	Unique	99.13 ± 0.04%
dataset 2	gsnap	Non-Unique	94.74 ± 0.14%
dataset 2	gsnap	Unique	99.25 ± 0.03%
dataset 2	rum	Non-Unique	96.73 ± 0.10%
dataset 2	rum	Unique	98.08 ± 0.03%
dataset 2	tophat	Non-Unique	93.63 ± 0.22%
dataset 2	tophat	Unique	97.94 ± 0.02%

* Note: Sites with a true coverage value less than 10 and true RDD level less than 10% per the simulated RNA-Seq dataset are not considered in this analysis.

Table 4.6 Sensitivity of RDD detection within RepeatMasker regions.

Dataset	Aligner	Repeat Masker	Sensitivity (%)
dataset 1	gsnap	In Repeat Masker Region	98.48 ± 0.11%
dataset 1	gsnap	Not In Repeat Masker Region	99.16 ± 0.02%
dataset 1	rum	In Repeat Masker Region	97.59 ± 0.16%
dataset 1	rum	Not In Repeat Masker Region	99.14 ± 0.02%
dataset 1	tophat2	In Repeat Masker Region	97.55 ± 0.13%
dataset 1	tophat2	Not In Repeat Masker Region	98.83 ± 0.01%
dataset 2	gsnap	In Repeat Masker Region	98.01 ± 0.12%
dataset 2	gsnap	Not In Repeat Masker Region	98.91 ± 0.02%
dataset 2	rum	In Repeat Masker Region	96.70 ± 0.18%
dataset 2	rum	Not In Repeat Masker Region	98.34 ± 0.08%
dataset 2	tophat2	In Repeat Masker Region	96.48 ± 0.27%
dataset 2	tophat2	Not In Repeat Masker Region	97.71 ± 0.04%

*Note: Sites with a true coverage value less than 10 and a true RDD level less than 10% per the simulated RNA-Seq dataset are not considered in this analysis.

Table 4.7 Sensitivity of RDD detection versus proximity to nearby RDDs.

Dataset	Aligner	Distance to Nearest RDD	Sensitivity (%)
dataset 1	gsnap	RDD not within 10 bp of another RDD	99.29 ± 0.01%
dataset 1	gsnap	RDD within 10 bp of another RDD	97.79 ± 0.04%
dataset 1	rum	RDD not within 10 bp of another RDD	99.35 ± 0.01%
dataset 1	rum	RDD within 10 bp of another RDD	96.75 ± 0.17%
dataset 1	tophat	RDD not within 10 bp of another RDD	99.06 ± 0.01%
dataset 1	tophat	RDD within 10 bp of another RDD	96.51 ± 0.11%
dataset 2	gsnap	RDD not within 10 bp of another RDD	99.06 ± 0.02%
dataset 2	gsnap	RDD within 10 bp of another RDD	95.66 ± 0.22%
dataset 2	rum	RDD not within 10 bp of another RDD	98.62 ± 0.01%
dataset 2	rum	RDD within 10 bp of another RDD	92.51 ± 0.12%
dataset 2	tophat	RDD not within 10 bp of another RDD	98.12 ± 0.02%
dataset 2	tophat	RDD within 10 bp of another RDD	91.75 ± 0.25%

*Note: Sites with coverage less than 10 and an RDD level less than 10% in the simulated dataset are not considered in this analysis.

4.3.4 Correlation between true versus observed RDD levels

In many studies, the mere detection of RDDs is not sufficient. For example, in studies on RNA editing or differential allelic expression, information about the degree or level of difference is important. Here we analyzed the correlation between true and observed RDD levels. Based on our previous analyses, we restricted our study to sites with a minimum coverage of 10x, minimum level of 10%, and minimum of 1 read bearing the sequence difference base. Using this threshold, we calculated the correlation between observed and true RDD levels to be relatively high, at approximately 98 ± 0.03% on average across all three replicates for all three aligners and both datasets (Figure 4.6 and Table 4.8). Although the true and observed levels correspond well, we found that roughly 20 to 40% of sites in each dataset for any aligner have observed levels that deviate from the true values by more than 5% (Figure 4.7). In particular, we found that in the majority (75 to 90%) of cases in which the observed and true levels deviate by at least 10%, the observed level underestimates the true level.

We hypothesized that one contributing factor to the discrepancy in RDD levels is the uniqueness or the ability of the region surrounding the site to be aligned accurately. Indeed, we found that approximately 27 to 34% of sites in which the true versus observed RDD levels differ by more than 30% are found in non-unique regions of the genome as determined by BLAT versus roughly 7 to 12% for those where the levels do not differ by 30% or more (Table 4.9).

Figure 4.6 True versus observed levels of RDDs. Here we plot the true or simulated level of RDD versus the observed level as determined by GSNAP, RUM, or Tophat for replicate 1. Sites with coverage less than 10 or an RDD level less than 10% per the simulated dataset are removed from consideration. Overall, we observed the correlation between true and observed levels to be approximately 98% in both datasets and across the various aligners and replicates.

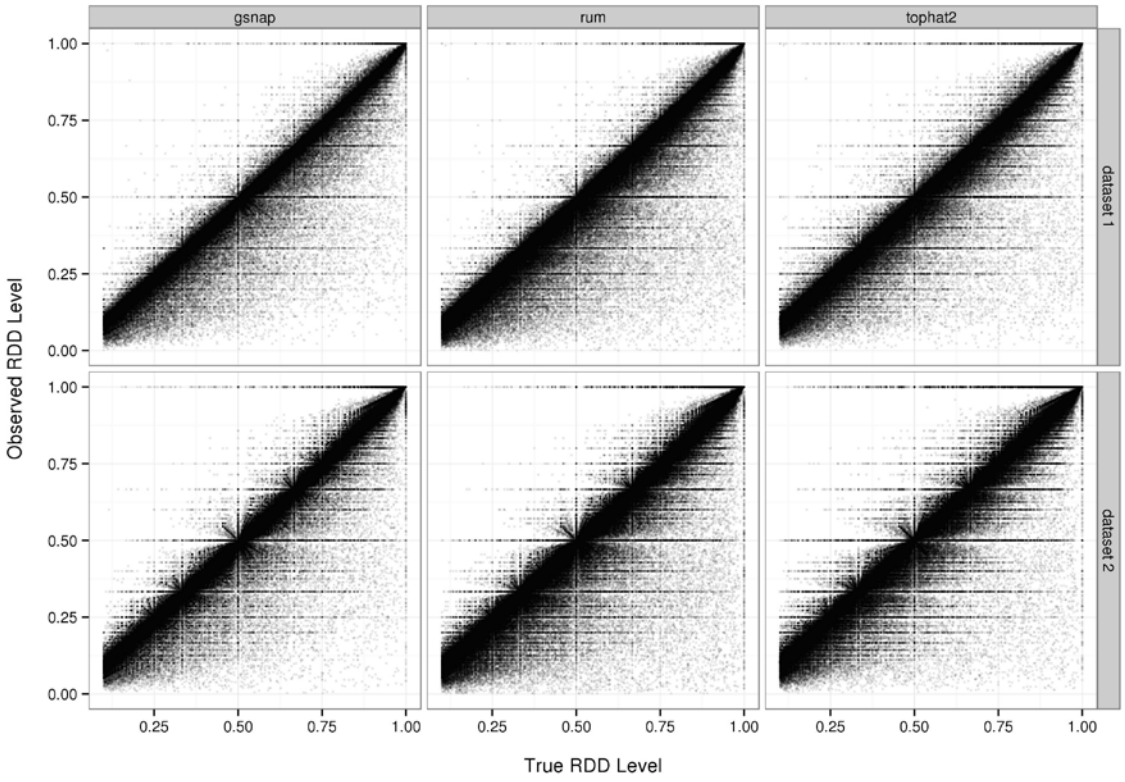


Table 4.8 Correlation between observed and true levels of RDDs.

Dataset	Aligner	Correlation (%)
dataset 1	gsnap	98.02 ± 0.03
dataset 1	rum	98.02 ± 0.03
dataset 1	tophat2	98.03 ± 0.03
dataset 2	gsnap	98.08 ± 0.03
dataset 2	rum	98.08 ± 0.03
dataset 2	tophat2	98.09 ± 0.03

*Note: Sites with coverage less than 10 and an RDD level less than 10% in the simulated dataset are not considered. Furthermore, sites must be identified as having at least 1 read bearing the sequence difference in the aligned dataset per GSNAP, RUM, or Tophat2 to be included in this analysis.

Figure 4.7 Percentage of sites with observed levels that deviate from true RDD level. Here we calculate the percentage of total sites in each dataset (*y*-axis) with observed levels that deviate from the true level of RDD by various amounts (*x*-axis). Colors refer to the aligner used for RDD detection.

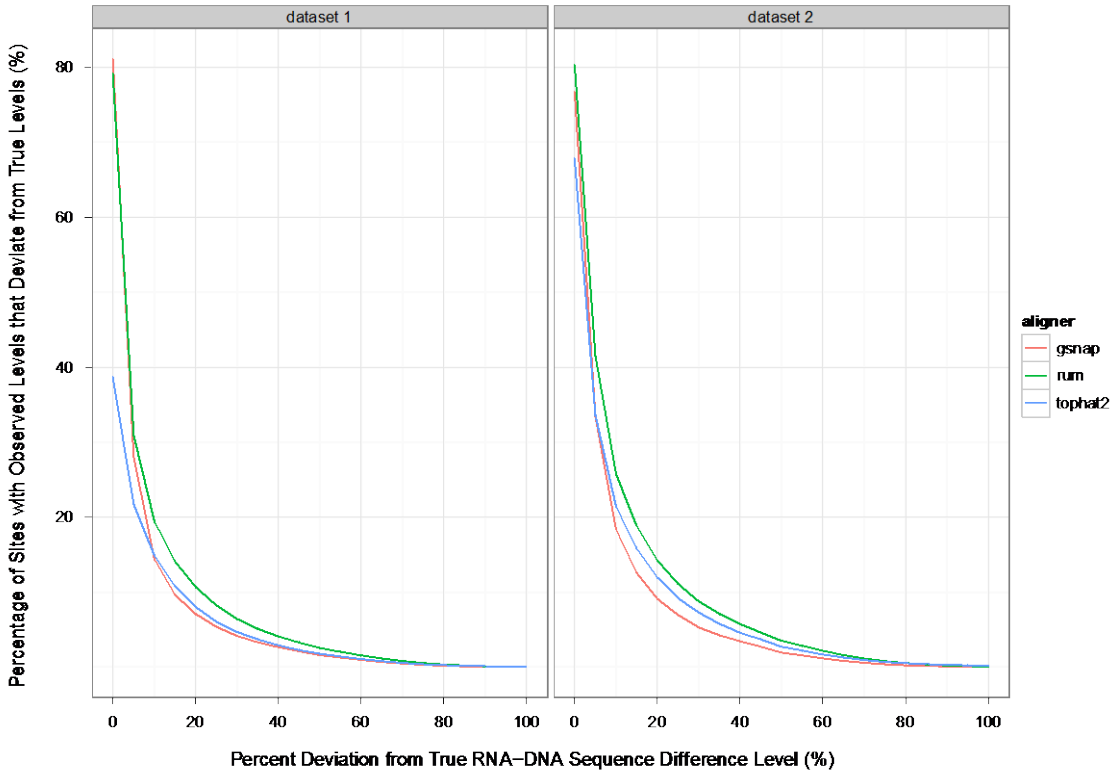


Table 4.9 Percent of sites where the observed and true levels deviate by more than 30% versus the uniqueness of the underlying site as determined by BLAT.

Dataset	Aligner	Levels differ by more than 30%	Percent of sites that are not unique as determined by BLAT
dataset 1	gsnap	TRUE	27.03 ± 0.56%
dataset 1	gsnap	FALSE	8.09 ± 0.03%
dataset 1	rum	TRUE	30.98 ± 0.52%
dataset 1	rum	FALSE	7.75 ± 0.01%
dataset 1	tophat	TRUE	31.55 ± 0.44%
dataset 1	tophat	FALSE	7.78 ± 0.02%
dataset 2	gsnap	TRUE	29.20 ± 0.57%
dataset 2	gsnap	FALSE	11.38 ± 0.05%
dataset 2	rum	TRUE	33.83 ± 0.42%
dataset 2	rum	FALSE	10.65 ± 0.07%
dataset 2	tophat	TRUE	29.17 ± 0.41%
dataset 2	tophat	FALSE	11.02 ± 0.06%

4.3.5 Receiver operating characteristic analysis of RDD detection

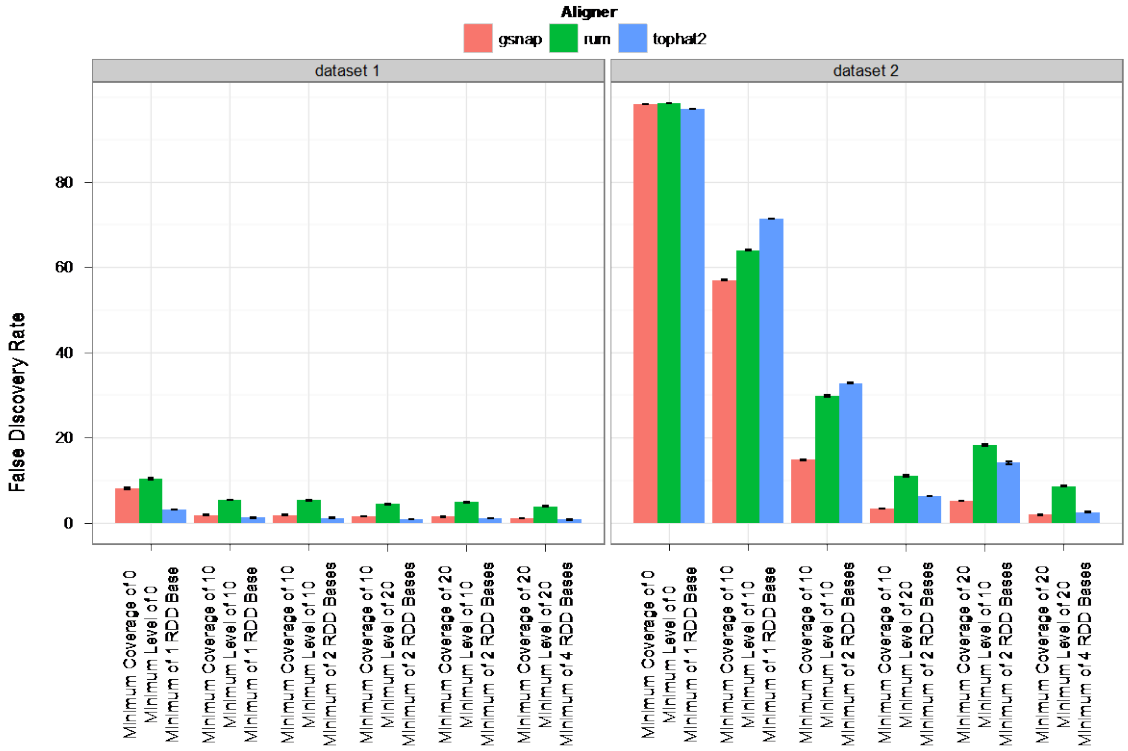
Next, we analyzed the false positive rate of RDD detection by evaluating the presence of differences at sites that were not simulated to represent RDDs. Using parameters we identified from our sensitivity analysis, we performed a receiver operating characteristic analysis on RDD detection genome-wide in each of the datasets (Table 4.10). Overall, we observed that using a 'minimum coverage of 10x, minimum level of 10%, and minimum of 1 read bearing the RDD base' cutoff, the false positive rate of sequence difference detection is low, averaging approximately 0.03% and 0.63% across the different aligners for datasets 1 and 2, respectively. However, given that the vast majority of sites in our datasets do not contain simulated sequence differences and are identified correctly as true negatives, these low false positives rates are not unexpected, as the false positive rate measures the percentage of true negatives that are incorrectly identified as having sequence differences. For a better understanding of the how false positives affect the analysis of RDDs, we evaluated the false discovery rate (FDR), or the percentage of sites identified as having sequence differences that in actuality are not true simulated differences. For dataset 1, we found the FDR to range from $1.31 \pm 4.06 \times 10^{-2}$ % in Tophat2 to $5.48 \pm 1.01 \times 10^{-1}$% in RUM when using a 'minimum coverage of 10x, minimum level of 10%, and minimum of 1 read bearing the RDD base' threshold. These relatively low false discovery rates indicate that in the absence of sequencing error, misalignment issues do not contribute significantly to the incidence of false positives. With the introduction of sequencing error in dataset 2, we found that the false discovery rates are much higher, ranging from approximately 57% in GSNAP to 71% in Tophat2. These results are not surprising, as our threshold requiring only 1 read bearing the RDD

base will result in false positives at sites with sequencing errors. With stricter thresholds on RDD detection, we found that the false discovery rate decreases dramatically. Requiring a minimum of 2 reads bearing the sequence difference in addition to a minimum coverage of 10 and level of 10% reduces the FDR by a least 50% for all three aligners in dataset 2. In the end, we found that with a ‘minimum coverage of 20x, minimum level of 20%, and minimum of 4 reads containing the sequence difference base’ cutoff, the FDR in dataset 2 ranges from approximately 2% in GSNAP to 9% in RUM (Figure 4.8).

Table 4.10 Receiver operating characteristic analysis of RDD detection.

Dataset	Aligner	Minimum coverage	Minimum level	Minimum RDD count	Accuracy (%)	True Positive Rate (%)	False Positive Rate (%)	False Discovery Rate (%)	Negative Predictive Value (%)
dataset 1	gsnap	0	0	1	99.90 ± 3.42E-03	98.36 ± 2.78E-02	0.09 ± 3.19E-03	8.20 ± 2.89E-01	99.98 ± 2.58E-04
dataset 1	gsnap	10	10	1	99.97 ± 6.97E-04	98.24 ± 1.66E-02	0.02 ± 5.41E-04	1.95 ± 6.00E-02	99.98 ± 2.20E-04
dataset 1	gsnap	10	10	2	99.97 ± 6.93E-04	98.20 ± 2.10E-02	0.02 ± 5.19E-04	1.92 ± 5.86E-02	99.98 ± 2.72E-04
dataset 1	gsnap	20	10	2	99.97 ± 7.30E-04	98.38 ± 3.13E-02	0.01 ± 4.44E-04	1.55 ± 4.67E-02	99.99 ± 2.93E-04
dataset 1	gsnap	10	20	2	99.96 ± 4.82E-04	97.02 ± 3.07E-02	0.01 ± 4.45E-04	1.62 ± 6.14E-02	99.98 ± 3.91E-04
dataset 1	gsnap	20	20	4	99.97 ± 5.13E-04	97.16 ± 3.45E-02	0.01 ± 3.50E-04	1.23 ± 4.69E-02	99.98 ± 4.23E-04
dataset 1	rum	0	0	1	99.87 ± 1.75E-03	97.76 ± 5.76E-02	0.11 ± 2.30E-03	10.46 ± 1.83E-01	99.98 ± 5.28E-04
dataset 1	rum	10	10	1	99.93 ± 4.30E-04	97.73 ± 2.52E-02	0.05 ± 6.72E-04	5.48 ± 1.01E-01	99.98 ± 3.60E-04
dataset 1	rum	10	10	2	99.93 ± 4.29E-04	97.68 ± 2.66E-02	0.05 ± 6.61E-04	5.45 ± 9.96E-02	99.98 ± 3.75E-04
dataset 1	rum	20	10	2	99.94 ± 6.64E-04	97.92 ± 2.94E-02	0.04 ± 8.25E-04	5.00 ± 1.09E-01	99.98 ± 3.85E-04
dataset 1	rum	10	20	2	99.94 ± 4.77E-04	96.21 ± 3.37E-02	0.04 ± 5.29E-04	4.51 ± 9.04E-02	99.97 ± 4.09E-04
dataset 1	rum	20	20	4	99.94 ± 5.48E-04	96.41 ± 4.60E-02	0.03 ± 5.74E-04	3.95 ± 9.40E-02	99.97 ± 5.28E-04
dataset 1	tophat2	0	0	1	99.86 ± 4.94E-04	88.50 ± 6.09E-02	0.03 ± 4.70E-04	3.18 ± 3.72E-02	99.89 ± 3.64E-04
dataset 1	tophat2	10	10	1	99.96 ± 4.09E-05	96.23 ± 4.05E-02	0.01 ± 3.51E-04	1.31 ± 4.06E-02	99.97 ± 3.11E-04
dataset 1	tophat2	10	10	2	99.95 ± 1.78E-04	95.86 ± 2.98E-02	0.01 ± 3.84E-04	1.26 ± 4.50E-02	99.96 ± 2.43E-04
dataset 1	tophat2	20	10	2	99.97 ± 4.50E-04	97.12 ± 6.48E-02	0.01 ± 2.72E-04	1.23 ± 3.59E-02	99.98 ± 6.38E-04
dataset 1	tophat2	10	20	2	99.96 ± 7.57E-04	95.01 ± 7.64E-02	0.01 ± 2.42E-04	0.88 ± 3.27E-02	99.96 ± 6.81E-04
dataset 1	tophat2	20	20	4	99.96 ± 6.68E-04	95.90 ± 8.61E-02	0.01 ± 1.73E-04	0.83 ± 2.64E-02	99.97 ± 7.01E-04
dataset 2	gsnap	0	0	1	94.12 ± 7.50E-04	97.69 ± 3.44E-02	5.89 ± 7.49E-04	98.38 ± 2.27E-03	100.00 ± 3.42E-05
dataset 2	gsnap	10	10	1	99.52 ± 8.19E-04	98.11 ± 1.48E-02	0.47 ± 8.59E-04	57.05 ± 6.61E-02	99.99 ± 4.57E-05
dataset 2	gsnap	10	10	2	99.93 ± 4.32E-04	97.96 ± 1.71E-02	0.06 ± 4.61E-04	14.95 ± 1.04E-01	99.99 ± 5.27E-05
dataset 2	gsnap	20	10	2	99.97 ± 4.12E-04	98.25 ± 2.70E-02	0.02 ± 4.86E-04	5.26 ± 9.22E-02	99.99 ± 7.58E-05
dataset 2	gsnap	10	20	2	99.98 ± 1.97E-04	96.98 ± 1.60E-02	0.01 ± 2.22E-04	3.46 ± 7.01E-02	99.99 ± 7.41E-05
dataset 2	gsnap	20	20	4	99.98 ± 4.87E-04	97.09 ± 3.57E-02	0.01 ± 4.43E-04	2.02 ± 1.19E-01	99.99 ± 1.41E-04
dataset 2	rum	0	0	1	93.74 ± 1.83E-03	96.52 ± 1.93E-02	6.26 ± 1.84E-03	98.54 ± 1.84E-03	100.00 ± 2.44E-05
dataset 2	rum	10	10	1	99.41 ± 2.26E-03	97.22 ± 5.03E-02	0.58 ± 2.11E-03	64.05 ± 1.00E-01	99.99 ± 1.66E-04
dataset 2	rum	10	10	2	99.85 ± 1.61E-03	97.03 ± 4.28E-02	0.14 ± 1.48E-03	29.87 ± 2.39E-01	99.99 ± 1.40E-04
dataset 2	rum	20	10	2	99.91 ± 1.41E-03	97.36 ± 6.64E-02	0.08 ± 1.17E-03	18.38 ± 2.07E-01	99.99 ± 2.47E-04
dataset 2	rum	10	20	2	99.95 ± 4.84E-04	95.55 ± 3.07E-02	0.04 ± 4.28E-04	11.12 ± 1.25E-01	99.99 ± 9.71E-05
dataset 2	rum	20	20	4	99.96 ± 6.51E-04	95.56 ± 3.96E-02	0.03 ± 5.40E-04	8.67 ± 1.34E-01	99.99 ± 1.70E-04
dataset 2	tophat2	0	0	1	95.13 ± 1.57E-03	81.63 ± 1.06E-01	4.84 ± 1.36E-03	97.29 ± 2.88E-03	99.97 ± 2.45E-04
dataset 2	tophat2	10	10	1	99.14 ± 3.84E-03	95.47 ± 5.23E-02	0.85 ± 3.57E-03	71.44 ± 5.99E-02	99.98 ± 3.14E-04
dataset 2	tophat2	10	10	2	99.82 ± 5.50E-04	94.83 ± 5.67E-02	0.17 ± 4.14E-04	32.94 ± 1.32E-01	99.98 ± 3.17E-04
dataset 2	tophat2	20	10	2	99.93 ± 3.87E-04	96.66 ± 1.04E-01	0.06 ± 7.42E-04	14.18 ± 2.98E-01	99.99 ± 4.50E-04
dataset 2	tophat2	10	20	2	99.96 ± 2.06E-04	93.97 ± 5.72E-02	0.02 ± 8.90E-05	6.43 ± 6.10E-02	99.98 ± 2.81E-04
dataset 2	tophat2	20	20	4	99.98 ± 4.98E-04	95.14 ± 7.79E-02	0.01 ± 1.70E-04	2.60 ± 6.07E-02	99.99 ± 4.07E-04

Figure 4.8 False discovery rate of RNA-DNA sequence difference detection. Here we depict the false discovery rate of RDD detection under various thresholds on the coverage, level of sequence difference, and number of reads bearing the sequence difference base per the aligner. Calculations are averaged across the three replicates and error bars represent standard deviation values. Colors refer to the aligner used to detect RDDs.



4.3.6 Evaluation of filters in removing false positive RDDs

Many previous studies on RNA editing and RDDs attempt to remove false positive sites using various filters (Bahn et al. 2012; Ju et al. 2011; Kleinman et al. 2012; Peng et al. 2012; Ramaswami et al. 2012). We investigated the effectiveness of some of these measures in eliminating false positives. The first filter we analyzed involves using BLAT to determine whether the sequence surrounding the RDD site can be aligned to other homologous regions of the genome (see Materials and Methods). Using a 'minimum coverage of 20x, minimum level of 20%, and minimum of 4 reads bearing the sequence difference' to identify RDDs, we observed that the BLAT method removes approximately 14% and 28% of false positives found by GSNAP in datasets 1 and 2 respectively, but only filters out roughly 1 to 5% of true positives in either datasets (Table 4.11). As expected, we observed that the performance of the BLAT filter varies depending on the repetitive nature of the underlying flanking sequence. For example, within RepeatMasker regions, approximately 22% of false positives and 19% of true positives are filtered out, whereas outside of RepeatMasker regions, roughly 14% of false positives are removed compared to less than 1% of true positives (Table 4.11). Interestingly, the difference between the percentages of false versus true positives removed by the BLAT method is largest for RUM, followed by GSNAP and Tophat2 (Figure 4.9). Overall, we found that the BLAT filtering approach decreased the FDR of RDD detection for GSNAP by approximately 13% in dataset 1 and 24% in dataset 2 (Figure 4.10; Table 4.12).

Table 4.11 Percentage of true versus false positive RDDs removed by BLAT filter.

Dataset	Aligner	Minimum Coverage	Minimum Level	Minimum RDD Count	Region	Percent of True Positives Removed	Percent of False Positives Removed
dataset 1	gsnap	10	10	2	Rmsk	18.59 ± 0.41	24.77 ± 4.61
dataset 1	gsnap	10	10	2	Not in Rmsk	0.25 ± 0.02	14.56 ± 1.30
dataset 1	gsnap	10	10	2	Total	1.55 ± 0.06	15.33 ± 1.56
dataset 1	rum	10	10	2	Rmsk	18.37 ± 0.17	24.65 ± 5.36
dataset 1	rum	10	10	2	Not in Rmsk	0.26 ± 0.02	27.04 ± 0.33
dataset 1	rum	10	10	2	Total	1.52 ± 0.04	26.83 ± 0.73
dataset 1	tophat	10	10	2	Rmsk	17.75 ± 0.54	15.49 ± 2.66
dataset 1	tophat	10	10	2	Not in Rmsk	0.23 ± 0.01	4.80 ± 0.34
dataset 1	tophat	10	10	2	Total	1.27 ± 0.01	5.52 ± 0.28
dataset 1	gsnap	10	20	2	Rmsk	18.53 ± 0.46	27.00 ± 4.25
dataset 1	gsnap	10	20	2	Not in Rmsk	0.25 ± 0.02	17.16 ± 1.34
dataset 1	gsnap	10	20	2	Total	1.54 ± 0.06	17.89 ± 1.54
dataset 1	rum	10	20	2	Rmsk	18.25 ± 0.20	25.79 ± 6.71
dataset 1	rum	10	20	2	Not in Rmsk	0.25 ± 0.01	32.94 ± 0.45
dataset 1	rum	10	20	2	Total	1.50 ± 0.04	32.36 ± 0.96
dataset 1	tophat	10	20	2	Rmsk	17.63 ± 0.49	14.58 ± 2.17
dataset 1	tophat	10	20	2	Not in Rmsk	0.23 ± 0.01	3.99 ± 0.33
dataset 1	tophat	10	20	2	Total	1.26 ± 0.01	4.75 ± 0.15
dataset 1	gsnap	20	10	2	Rmsk	18.84 ± 0.53	21.18 ± 4.18
dataset 1	gsnap	20	10	2	Not in Rmsk	0.24 ± 0.02	11.42 ± 1.65
dataset 1	gsnap	20	10	2	Total	1.47 ± 0.07	12.09 ± 1.84
dataset 1	rum	20	10	2	Rmsk	18.30 ± 0.25	22.56 ± 6.05
dataset 1	rum	20	10	2	Not in Rmsk	0.25 ± 0.01	24.80 ± 0.41
dataset 1	rum	20	10	2	Total	1.41 ± 0.03	24.60 ± 0.70
dataset 1	tophat	20	10	2	Rmsk	16.29 ± 0.60	15.91 ± 3.32
dataset 1	tophat	20	10	2	Not in Rmsk	0.23 ± 0.01	4.55 ± 0.52
dataset 1	tophat	20	10	2	Total	1.08 ± 0.00	5.31 ± 0.27
dataset 1	gsnap	20	20	4	Rmsk	18.75 ± 0.62	21.96 ± 4.78
dataset 1	gsnap	20	20	4	Not in Rmsk	0.23 ± 0.01	13.56 ± 1.72
dataset 1	gsnap	20	20	4	Total	1.46 ± 0.07	14.11 ± 1.94
dataset 1	rum	20	20	4	Rmsk	18.16 ± 0.22	23.57 ± 7.54
dataset 1	rum	20	20	4	Not in Rmsk	0.24 ± 0.01	30.90 ± 0.54
dataset 1	rum	20	20	4	Total	1.39 ± 0.03	30.31 ± 1.07
dataset 1	tophat	20	20	4	Rmsk	16.03 ± 0.47	14.81 ± 2.89
dataset 1	tophat	20	20	4	Not in Rmsk	0.23 ± 0.01	3.73 ± 0.47
dataset 1	tophat	20	20	4	Total	1.06 ± 0.01	4.52 ± 0.19
dataset 2	gsnap	10	10	2	Rmsk	24.77 ± 0.47	25.51 ± 0.30
dataset 2	gsnap	10	10	2	Not in Rmsk	0.22 ± 0.02	5.87 ± 0.18
dataset 2	gsnap	10	10	2	Total	6.43 ± 0.07	12.40 ± 0.20
dataset 2	rum	10	10	2	Rmsk	23.18 ± 0.31	35.28 ± 0.25
dataset 2	rum	10	10	2	Not in Rmsk	0.23 ± 0.02	14.73 ± 0.09
dataset 2	rum	10	10	2	Total	5.86 ± 0.05	21.84 ± 0.13
dataset 2	tophat	10	10	2	Rmsk	22.09 ± 0.29	23.39 ± 0.70
dataset 2	tophat	10	10	2	Not in Rmsk	0.23 ± 0.02	0.73 ± 0.02
dataset 2	tophat	10	10	2	Total	2.91 ± 0.02	4.04 ± 0.08
dataset 2	gsnap	10	20	2	Rmsk	24.78 ± 0.39	27.37 ± 1.06
dataset 2	gsnap	10	20	2	Not in Rmsk	0.21 ± 0.01	23.15 ± 0.43
dataset 2	gsnap	10	20	2	Total	6.44 ± 0.05	24.23 ± 0.53
dataset 2	rum	10	20	2	Rmsk	23.11 ± 0.25	46.08 ± 0.15
dataset 2	rum	10	20	2	Not in Rmsk	0.22 ± 0.02	35.67 ± 0.42
dataset 2	rum	10	20	2	Total	5.86 ± 0.04	38.71 ± 0.30
dataset 2	tophat	10	20	2	Rmsk	22.00 ± 0.22	25.09 ± 2.05
dataset 2	tophat	10	20	2	Not in Rmsk	0.22 ± 0.02	2.24 ± 0.17
dataset 2	tophat	10	20	2	Total	2.89 ± 0.03	6.26 ± 0.51

Dataset	Aligner	Minimum Coverage	Minimum Level	Minimum RDD Count	Region	Percent of True Positives Removed	Percent of False Positives Removed
dataset 2	gsnap	20	10	2	Rmsk	24.25 ± 0.61	25.81 ± 0.79
dataset 2	gsnap	20	10	2	Not in Rmsk	0.23 ± 0.02	12.52 ± 0.77
dataset 2	gsnap	20	10	2	Total	4.84 ± 0.09	16.00 ± 0.68
dataset 2	rum	20	10	2	Rmsk	22.34 ± 0.50	37.67 ± 0.76
dataset 2	rum	20	10	2	Not in Rmsk	0.24 ± 0.02	25.60 ± 0.09
dataset 2	rum	20	10	2	Total	4.32 ± 0.07	28.97 ± 0.20
dataset 2	tophat	20	10	2	Rmsk	19.57 ± 1.07	22.03 ± 1.72
dataset 2	tophat	20	10	2	Not in Rmsk	0.25 ± 0.02	1.68 ± 0.07
dataset 2	tophat	20	10	2	Total	1.96 ± 0.10	4.08 ± 0.16
dataset 2	gsnap	20	20	4	Rmsk	24.27 ± 0.48	28.57 ± 0.85
dataset 2	gsnap	20	20	4	Not in Rmsk	0.22 ± 0.01	27.55 ± 1.05
dataset 2	gsnap	20	20	4	Total	4.83 ± 0.06	27.74 ± 0.72
dataset 2	rum	20	20	4	Rmsk	22.26 ± 0.40	43.83 ± 1.33
dataset 2	rum	20	20	4	Not in Rmsk	0.23 ± 0.02	42.13 ± 0.34
dataset 2	rum	20	20	4	Total	4.29 ± 0.06	42.51 ± 0.25
dataset 2	tophat	20	20	4	Rmsk	19.56 ± 1.18	23.88 ± 5.57
dataset 2	tophat	20	20	4	Not in Rmsk	0.23 ± 0.01	5.28 ± 0.46
dataset 2	tophat	20	20	4	Total	1.94 ± 0.10	8.15 ± 1.27

Figure 4.9a Percentage of false versus true positive RDDs removed by BLAT filter for dataset 1. Here we depict the percentage of false positives versus true positives that are removed when using the BLAT filter for dataset 1. Colors refer to the aligner used for RDD detection.

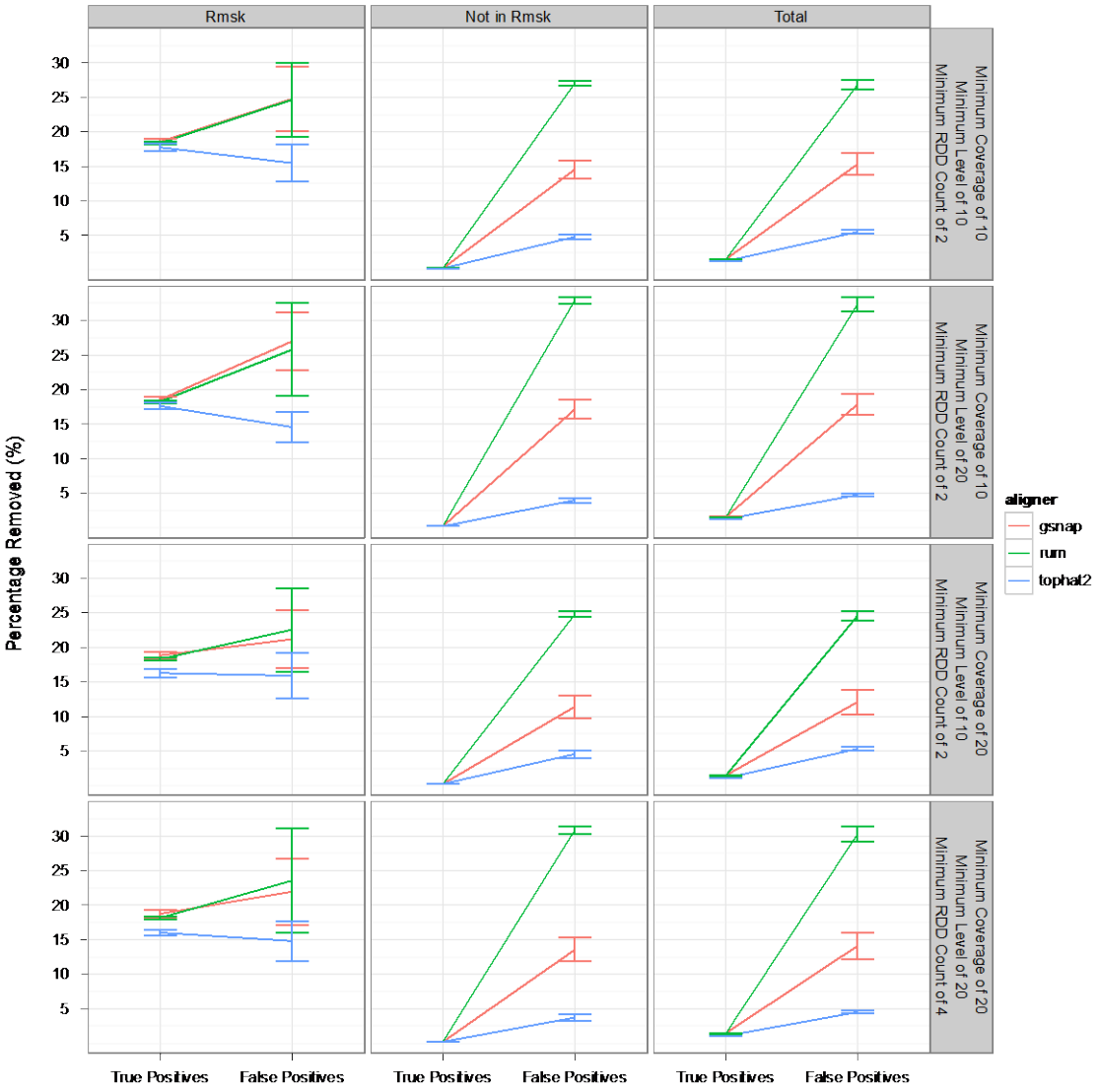


Figure 4.9b Percentage of false versus true positive RDDs removed by BLAT filter for dataset 2. Here we depict the percentage of false positives versus true positives that are removed when using the BLAT filter for dataset 2. Colors refer to the aligner used for RDD detection.

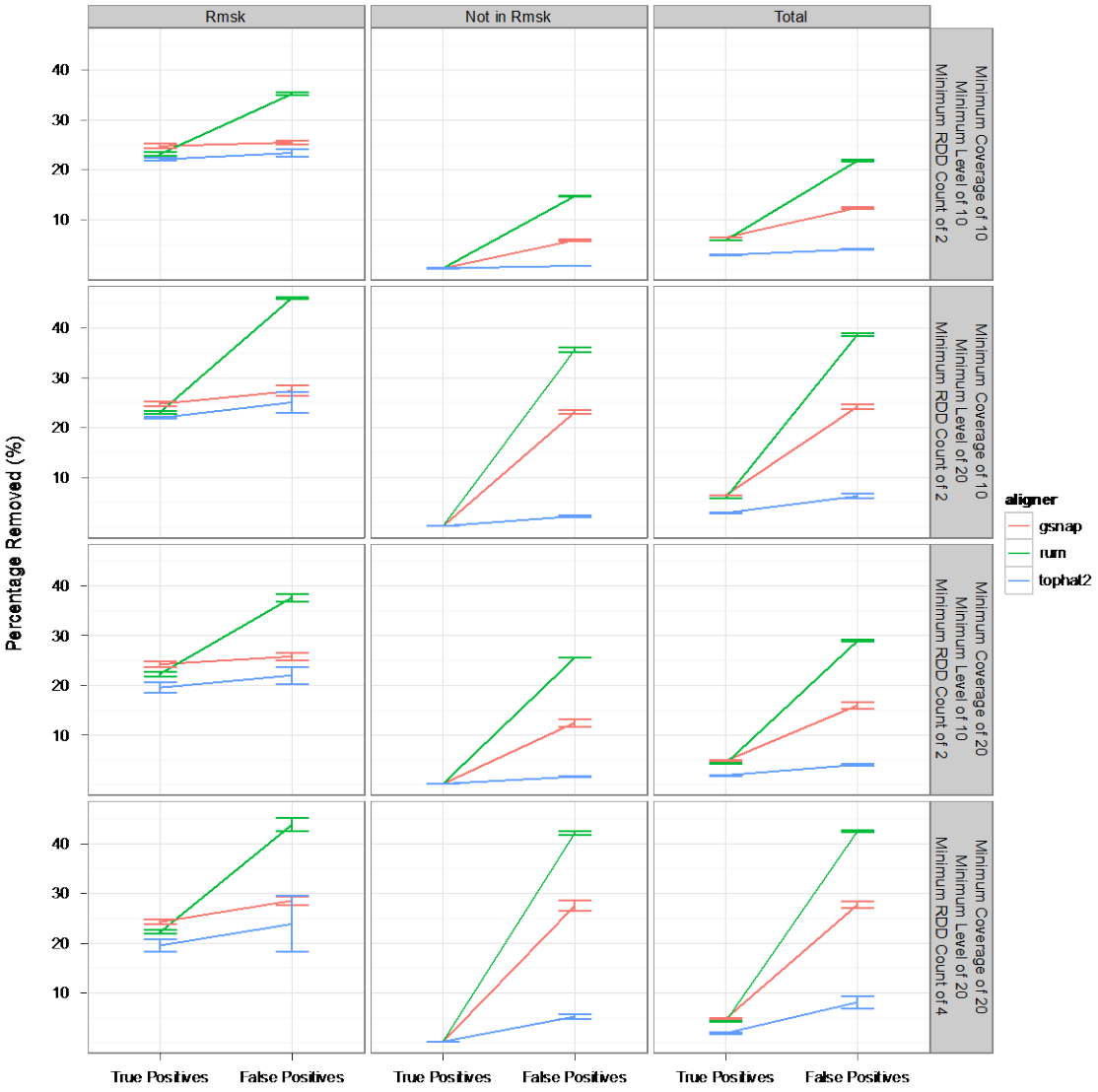


Figure 4.10 Effect of BLAT filter on false discovery rate of RDD detection. Here we depict the effect of the BLAT filter on the FDR for various aligners and thresholds for identification of sequence differences. Colors refer to the aligner used for RDD detection.

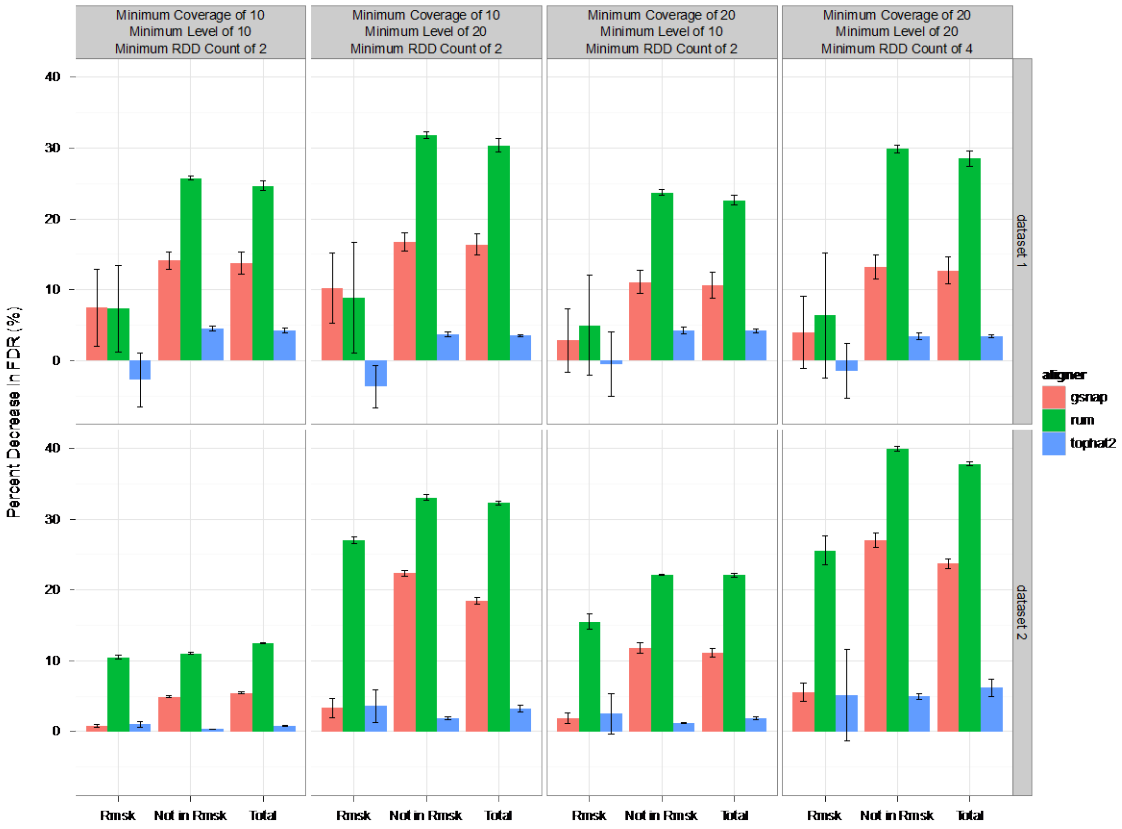


Table 4.12 Effect of BLAT filter on false discovery rate of RDD detection.

Dataset	Aligner	Minimum Coverage	Minimum Level	Minimum RDD Count	Region	FDR Before BLAT Filter	FDR After BLAT Filter	Percent Decrease in FDR
dataset 1	gsnap	10	10	2	Rmsk	2.02 ± 0.18	1.86 ± 0.10	7.46 ± 5.45
dataset 1	gsnap	10	10	2	Not in Rmsk	1.92 ± 0.05	1.65 ± 0.02	14.11 ± 1.27
dataset 1	gsnap	10	10	2	Total	1.92 ± 0.06	1.66 ± 0.02	13.76 ± 1.52
dataset 1	rum	10	10	2	Rmsk	6.06 ± 0.71	5.64 ± 1.00	7.31 ± 6.14
dataset 1	rum	10	10	2	Not in Rmsk	5.41 ± 0.06	4.01 ± 0.06	25.77 ± 0.33
dataset 1	rum	10	10	2	Total	5.45 ± 0.10	4.11 ± 0.11	24.65 ± 0.71
dataset 1	tophat2	10	10	2	Rmsk	1.43 ± 0.17	1.48 ± 0.21	-2.73 ± 3.84
dataset 1	tophat2	10	10	2	Not in Rmsk	1.24 ± 0.04	1.19 ± 0.04	4.52 ± 0.33
dataset 1	tophat2	10	10	2	Total	1.26 ± 0.04	1.20 ± 0.05	4.25 ± 0.29
dataset 1	gsnap	10	20	2	Rmsk	1.68 ± 0.20	1.50 ± 0.13	10.25 ± 4.92
dataset 1	gsnap	10	20	2	Not in Rmsk	1.61 ± 0.05	1.34 ± 0.04	16.72 ± 1.31
dataset 1	gsnap	10	20	2	Total	1.61 ± 0.06	1.35 ± 0.04	16.38 ± 1.51
dataset 1	rum	10	20	2	Rmsk	4.97 ± 0.73	4.56 ± 1.01	8.86 ± 7.76
dataset 1	rum	10	20	2	Not in Rmsk	4.47 ± 0.06	3.05 ± 0.06	31.78 ± 0.46
dataset 1	rum	10	20	2	Total	4.50 ± 0.09	3.14 ± 0.11	30.35 ± 0.95
dataset 1	tophat2	10	20	2	Rmsk	1.05 ± 0.13	1.09 ± 0.13	-3.67 ± 3.03
dataset 1	tophat2	10	20	2	Not in Rmsk	0.87 ± 0.03	0.83 ± 0.03	3.74 ± 0.32
dataset 1	tophat2	10	20	2	Total	0.88 ± 0.03	0.85 ± 0.03	3.50 ± 0.15
dataset 1	gsnap	20	10	2	Rmsk	1.58 ± 0.12	1.53 ± 0.07	2.86 ± 4.50
dataset 1	gsnap	20	10	2	Not in Rmsk	1.55 ± 0.04	1.38 ± 0.02	11.06 ± 1.61
dataset 1	gsnap	20	10	2	Total	1.55 ± 0.05	1.39 ± 0.01	10.62 ± 1.80
dataset 1	rum	20	10	2	Rmsk	5.82 ± 0.84	5.56 ± 1.17	4.97 ± 7.08
dataset 1	rum	20	10	2	Not in Rmsk	4.95 ± 0.06	3.77 ± 0.06	23.68 ± 0.39
dataset 1	rum	20	10	2	Total	5.00 ± 0.11	3.87 ± 0.12	22.61 ± 0.69
dataset 1	tophat2	20	10	2	Rmsk	1.52 ± 0.21	1.52 ± 0.21	-0.47 ± 4.57
dataset 1	tophat2	20	10	2	Not in Rmsk	1.21 ± 0.03	1.16 ± 0.03	4.28 ± 0.51
dataset 1	tophat2	20	10	2	Total	1.23 ± 0.04	1.18 ± 0.04	4.22 ± 0.27
dataset 1	gsnap	20	20	4	Rmsk	1.20 ± 0.10	1.15 ± 0.06	3.93 ± 5.10
dataset 1	gsnap	20	20	4	Not in Rmsk	1.23 ± 0.04	1.07 ± 0.02	13.22 ± 1.70
dataset 1	gsnap	20	20	4	Total	1.23 ± 0.05	1.07 ± 0.03	12.70 ± 1.89
dataset 1	rum	20	20	4	Rmsk	4.54 ± 0.86	4.29 ± 1.19	6.40 ± 8.84
dataset 1	rum	20	20	4	Not in Rmsk	3.90 ± 0.06	2.74 ± 0.05	29.89 ± 0.52
dataset 1	rum	20	20	4	Total	3.95 ± 0.09	2.82 ± 0.11	28.49 ± 1.06
dataset 1	tophat2	20	20	4	Rmsk	1.11 ± 0.13	1.13 ± 0.13	-1.44 ± 3.84
dataset 1	tophat2	20	20	4	Not in Rmsk	0.82 ± 0.02	0.79 ± 0.02	3.48 ± 0.46

Dataset	Aligner	Minimum Coverage	Minimum Level	Minimum RDD Count	Region	FDR Before BLAT Filter	FDR After BLAT Filter	Percent Decrease in FDR
dataset 1	tophat2	20	20	4	Total	0.83 ± 0.03	0.80 ± 0.03	3.47 ± 0.20
dataset 2	gsnap	10	10	2	Rmsk	18.77 ± 0.17	18.62 ± 0.20	0.79 ± 0.19
dataset 2	gsnap	10	10	2	Not in Rmsk	13.57 ± 0.13	12.90 ± 0.12	4.93 ± 0.16
dataset 2	gsnap	10	10	2	Total	14.94 ± 0.10	14.13 ± 0.10	5.48 ± 0.14
dataset 2	rum	10	10	2	Rmsk	37.54 ± 0.27	33.62 ± 0.29	10.45 ± 0.28
dataset 2	rum	10	10	2	Not in Rmsk	26.95 ± 0.26	23.98 ± 0.26	11.05 ± 0.11
dataset 2	rum	10	10	2	Total	29.87 ± 0.24	26.13 ± 0.21	12.54 ± 0.07
dataset 2	tophat2	10	10	2	Rmsk	36.97 ± 0.07	36.58 ± 0.10	1.06 ± 0.41
dataset 2	tophat2	10	10	2	Not in Rmsk	32.34 ± 0.15	32.23 ± 0.15	0.34 ± 0.00
dataset 2	tophat2	10	10	2	Total	32.94 ± 0.13	32.68 ± 0.12	0.78 ± 0.05
dataset 2	gsnap	10	20	2	Rmsk	3.48 ± 0.05	3.36 ± 0.01	3.33 ± 1.38
dataset 2	gsnap	10	20	2	Not in Rmsk	3.45 ± 0.08	2.68 ± 0.05	22.37 ± 0.41
dataset 2	gsnap	10	20	2	Total	3.46 ± 0.07	2.82 ± 0.04	18.48 ± 0.51
dataset 2	rum	10	20	2	Rmsk	12.91 ± 0.33	9.42 ± 0.29	27.06 ± 0.46
dataset 2	rum	10	20	2	Not in Rmsk	10.48 ± 0.09	7.02 ± 0.08	33.03 ± 0.42
dataset 2	rum	10	20	2	Total	11.09 ± 0.13	7.51 ± 0.08	32.28 ± 0.28
dataset 2	tophat2	10	20	2	Rmsk	8.94 ± 0.09	8.62 ± 0.14	3.62 ± 2.34
dataset 2	tophat2	10	20	2	Not in Rmsk	6.04 ± 0.06	5.92 ± 0.05	1.90 ± 0.15
dataset 2	tophat2	10	20	2	Total	6.40 ± 0.06	6.19 ± 0.03	3.25 ± 0.47
dataset 2	gsnap	20	10	2	Rmsk	7.04 ± 0.18	6.90 ± 0.19	1.92 ± 0.74
dataset 2	gsnap	20	10	2	Not in Rmsk	4.83 ± 0.08	4.26 ± 0.05	11.80 ± 0.74
dataset 2	gsnap	20	10	2	Total	5.26 ± 0.09	4.67 ± 0.06	11.18 ± 0.60
dataset 2	rum	20	10	2	Rmsk	25.42 ± 0.49	21.48 ± 0.68	15.50 ± 1.11
dataset 2	rum	20	10	2	Not in Rmsk	16.60 ± 0.21	12.93 ± 0.17	22.13 ± 0.11
dataset 2	rum	20	10	2	Total	18.38 ± 0.21	14.33 ± 0.20	22.07 ± 0.24
dataset 2	tophat2	20	10	2	Rmsk	18.00 ± 0.74	17.56 ± 1.17	2.52 ± 2.83
dataset 2	tophat2	20	10	2	Not in Rmsk	13.79 ± 0.28	13.62 ± 0.28	1.24 ± 0.05
dataset 2	tophat2	20	10	2	Total	14.18 ± 0.30	13.92 ± 0.31	1.86 ± 0.22
dataset 2	gsnap	20	20	4	Rmsk	1.98 ± 0.16	1.88 ± 0.17	5.57 ± 1.32
dataset 2	gsnap	20	20	4	Not in Rmsk	2.02 ± 0.12	1.48 ± 0.07	26.99 ± 1.02
dataset 2	gsnap	20	20	4	Total	2.02 ± 0.12	1.54 ± 0.08	23.70 ± 0.67
dataset 2	rum	20	20	4	Rmsk	10.65 ± 0.54	7.93 ± 0.62	25.55 ± 2.02
dataset 2	rum	20	20	4	Not in Rmsk	8.18 ± 0.10	4.92 ± 0.08	39.93 ± 0.36
dataset 2	rum	20	20	4	Total	8.65 ± 0.13	5.38 ± 0.11	37.79 ± 0.31
dataset 2	tophat2	20	20	4	Rmsk	4.40 ± 0.30	4.17 ± 0.34	5.16 ± 6.49
dataset 2	tophat2	20	20	4	Not in Rmsk	2.39 ± 0.05	2.27 ± 0.05	4.95 ± 0.45
dataset 2	tophat2	20	20	4	Total	2.57 ± 0.06	2.41 ± 0.07	6.19 ± 1.25

Pseudogenes are non-functioning homologs of genes that are either not expressed or unable to be translated into protein product, and their high sequence similarity to functioning genes can result in false positive sequence difference calls. We observed that the removal of all sequence differences lying within pseudogenes as annotated by Gencode version 13 (Harrow et al. 2006) decreases the FDR of RDD detection using GSNAP by approximately 45 to 50% in both datasets (Table 4.13).

Misalignments near exon-exon junctions can commonly lead to the identification of false positive sequence differences. We evaluated the effect of such incorrectly spliced alignments on sequence difference detection and found that roughly 3% of the false positives identified by GSNAP in dataset 1 and 5% of those found in dataset 2 are in intronic sequences within 6 bp of exon-exon junctions. Removal of all sites in introns within 6 bp of splice junctions leads to a roughly 2 to 4% decrease in the false discovery rate for GSNAP. The other two aligners, RUM and Tophat2, are more robust to misalignments near splice junctions, as less than 1 to 2% of false positives detected by either aligner are in introns near exon-exon junctions (Table 4.14).

Table 4.13 Effect of removing RDDs in pseudogenes on the false discovery rate of sequence difference detection.

Dataset	Aligner	Minimum Coverage	Minimum Level	Minimum RDD Count	FDR Before Pseudogene Filter (%)	FDR After Pseudogene Filter (%)	Percent Decrease in FDR
dataset 1	gsnap	10	10	2	1.92 ± 0.06	0.95 ± 0.05	50.72 ± 1.41
dataset 1	rum	10	10	2	5.45 ± 0.10	2.35 ± 0.01	56.92 ± 0.94
dataset 1	tophat2	10	10	2	1.26 ± 0.04	0.85 ± 0.03	32.50 ± 1.62
dataset 1	gsnap	10	20	2	1.61 ± 0.06	0.74 ± 0.04	53.89 ± 1.14
dataset 1	rum	10	20	2	4.50 ± 0.09	1.67 ± 0.01	62.98 ± 0.80
dataset 1	tophat2	10	20	2	0.88 ± 0.03	0.57 ± 0.02	35.23 ± 1.07
dataset 1	gsnap	20	10	2	1.55 ± 0.05	0.79 ± 0.05	49.18 ± 1.79
dataset 1	rum	20	10	2	5.00 ± 0.11	2.24 ± 0.05	55.27 ± 1.86
dataset 1	tophat2	20	10	2	1.23 ± 0.04	0.87 ± 0.04	28.83 ± 1.74
dataset 1	gsnap	20	20	4	1.23 ± 0.05	0.58 ± 0.04	52.47 ± 1.73
dataset 1	rum	20	20	4	3.95 ± 0.09	1.50 ± 0.03	62.01 ± 1.70
dataset 1	tophat2	20	20	4	0.83 ± 0.03	0.58 ± 0.03	30.62 ± 1.48
dataset 2	gsnap	10	10	2	14.94 ± 0.10	13.90 ± 0.11	6.99 ± 0.16
dataset 2	rum	10	10	2	29.87 ± 0.24	25.21 ± 0.28	15.61 ± 0.26
dataset 2	tophat2	10	10	2	32.94 ± 0.13	32.62 ± 0.14	0.96 ± 0.03
dataset 2	gsnap	10	20	2	3.46 ± 0.07	2.39 ± 0.06	30.91 ± 0.39
dataset 2	rum	10	20	2	11.09 ± 0.13	7.39 ± 0.16	33.40 ± 0.75
dataset 2	tophat2	10	20	2	6.40 ± 0.06	5.98 ± 0.06	6.58 ± 0.17
dataset 2	gsnap	20	10	2	5.26 ± 0.09	4.19 ± 0.05	20.45 ± 0.96
dataset 2	rum	20	10	2	18.38 ± 0.21	13.44 ± 0.21	26.89 ± 0.41
dataset 2	tophat2	20	10	2	14.18 ± 0.30	13.62 ± 0.36	3.97 ± 0.53
dataset 2	gsnap	20	20	4	2.02 ± 0.12	1.10 ± 0.05	45.32 ± 1.81
dataset 2	rum	20	20	4	8.65 ± 0.13	5.04 ± 0.14	41.72 ± 0.70
dataset 2	tophat2	20	20	4	2.57 ± 0.06	2.08 ± 0.06	19.01 ± 1.00

Table 4.14 Effect of removing RDDs near exon junctions on the false discovery rate of sequence difference detection.

Dataset	Aligner	Minimum Coverage	Minimum Level	Minimum RDD Count	Percent of false positives near junctions	FDR before filter (%)	FDR after filter (%)	Percent FDR decrease
dataset 1	gsnap	10	10	2	2.64 ± 0.19	1.92 ± 0.06	1.88 ± 0.06	2.34 ± 0.19
dataset 1	rum	10	10	2	1.12 ± 0.03	5.45 ± 0.10	5.41 ± 0.10	0.82 ± 0.02
dataset 1	tophat2	10	10	2	0.55 ± 0.14	1.26 ± 0.04	1.25 ± 0.04	0.24 ± 0.15
dataset 1	gsnap	10	20	2	3.20 ± 0.24	1.61 ± 0.06	1.57 ± 0.06	2.90 ± 0.24
dataset 1	rum	10	20	2	1.35 ± 0.04	4.50 ± 0.09	4.46 ± 0.09	1.05 ± 0.04
dataset 1	tophat2	10	20	2	0.55 ± 0.18	0.88 ± 0.03	0.87 ± 0.03	0.24 ± 0.18
dataset 1	gsnap	20	10	2	1.85 ± 0.19	1.55 ± 0.05	1.53 ± 0.05	1.56 ± 0.19
dataset 1	rum	20	10	2	0.72 ± 0.09	5.00 ± 0.11	4.98 ± 0.11	0.43 ± 0.08
dataset 1	tophat2	20	10	2	0.66 ± 0.06	1.23 ± 0.04	1.22 ± 0.04	0.32 ± 0.07
dataset 1	gsnap	20	20	4	2.29 ± 0.23	1.23 ± 0.05	1.20 ± 0.05	2.01 ± 0.22
dataset 1	rum	20	20	4	0.83 ± 0.12	3.95 ± 0.09	3.92 ± 0.10	0.54 ± 0.11
dataset 1	tophat2	20	20	4	0.61 ± 0.07	0.83 ± 0.03	0.83 ± 0.03	0.28 ± 0.06
dataset 2	gsnap	10	10	2	1.47 ± 0.06	14.94 ± 0.10	14.80 ± 0.11	0.98 ± 0.06
dataset 2	rum	10	10	2	1.20 ± 0.06	29.87 ± 0.24	29.69 ± 0.23	0.62 ± 0.04
dataset 2	tophat2	10	10	2	0.41 ± 0.03	32.94 ± 0.13	32.93 ± 0.14	0.03 ± 0.03
dataset 2	gsnap	10	20	2	4.50 ± 0.24	3.46 ± 0.07	3.32 ± 0.08	4.06 ± 0.23
dataset 2	rum	10	20	2	2.53 ± 0.15	11.09 ± 0.13	10.87 ± 0.11	1.97 ± 0.14
dataset 2	tophat2	10	20	2	0.57 ± 0.14	6.40 ± 0.06	6.39 ± 0.05	0.19 ± 0.15
dataset 2	gsnap	20	10	2	2.42 ± 0.11	5.26 ± 0.09	5.16 ± 0.09	1.98 ± 0.09
dataset 2	rum	20	10	2	1.39 ± 0.07	18.38 ± 0.21	18.23 ± 0.20	0.86 ± 0.06
dataset 2	tophat2	20	10	2	0.50 ± 0.04	14.18 ± 0.30	14.16 ± 0.29	0.11 ± 0.03
dataset 2	gsnap	20	20	4	4.47 ± 0.46	2.02 ± 0.12	1.93 ± 0.12	4.08 ± 0.43
dataset 2	rum	20	20	4	2.02 ± 0.12	8.65 ± 0.13	8.51 ± 0.12	1.54 ± 0.12
dataset 2	tophat2	20	20	4	0.84 ± 0.61	2.57 ± 0.06	2.56 ± 0.06	0.46 ± 0.60

Lastly, we analyzed the effect of implementing the various bioinformatics filters altogether on eliminating false positive RDDs. In analyzing the RDDs obtained using the 'minimum coverage of 20x, minimum level of 20%, and minimum of 4 reads containing the sequence difference allele', we observed that the BLAT filtering method, removal of differences in pseudogenes, and elimination of intronic sites within 6 bp of exon junctions in combination removed nearly 30 to 70% of false positives depending on the aligner versus roughly 3 to 7% of true positives (Table 4.15). Overall, these various filters led to a decrease of approximately 50 to 70% in the FDR of RDD detection for GSNAP and RUM and a decrease of roughly 30% for Tophat2 (Table 4.16). While these various filters are successful in targeting false versus true positives for removal, a sizeable percentage of false positives remain.

Table 4.15 Percentage of true versus false positives removed by BLAT filter, pseudogene filter, and removal of intronic sites within 6 bp of exon junctions.

Dataset	Aligner	Minimum Coverage	Minimum Level	Minimum RDD Count	Region	Percent of True Positives Removed	Percent of False Positives Removed
dataset 1	gsnap	10	10	2	Rmsk	21.78 ± 0.45	53.56 ± 4.58
dataset 1	gsnap	10	10	2	Not in Rmsk	2.16 ± 0.03	60.37 ± 0.44
dataset 1	gsnap	10	10	2	Total	3.55 ± 0.08	59.88 ± 0.68
dataset 1	rum	10	10	2	Rmsk	21.79 ± 0.22	65.95 ± 6.68
dataset 1	rum	10	10	2	Not in Rmsk	2.25 ± 0.04	66.58 ± 0.44
dataset 1	rum	10	10	2	Total	3.61 ± 0.08	66.56 ± 0.56
dataset 1	tophat2	10	10	2	Rmsk	20.75 ± 0.53	47.07 ± 5.00
dataset 1	tophat2	10	10	2	Not in Rmsk	1.83 ± 0.01	37.22 ± 1.19
dataset 1	tophat2	10	10	2	Total	2.95 ± 0.03	37.91 ± 1.34
dataset 1	gsnap	10	20	2	Rmsk	21.70 ± 0.49	55.45 ± 3.15
dataset 1	gsnap	10	20	2	Not in Rmsk	2.11 ± 0.03	64.54 ± 0.39
dataset 1	gsnap	10	20	2	Total	3.50 ± 0.09	63.88 ± 0.34
dataset 1	rum	10	20	2	Rmsk	21.64 ± 0.26	68.17 ± 5.38
dataset 1	rum	10	20	2	Not in Rmsk	2.15 ± 0.04	72.45 ± 0.52
dataset 1	rum	10	20	2	Total	3.50 ± 0.08	72.15 ± 0.45
dataset 1	tophat2	10	20	2	Rmsk	20.58 ± 0.60	45.62 ± 2.05
dataset 1	tophat2	10	20	2	Not in Rmsk	1.79 ± 0.01	39.18 ± 0.98
dataset 1	tophat2	10	20	2	Total	2.90 ± 0.03	39.64 ± 1.02
dataset 1	gsnap	20	10	2	Rmsk	21.85 ± 0.56	50.33 ± 5.77
dataset 1	gsnap	20	10	2	Not in Rmsk	1.97 ± 0.05	56.58 ± 0.63
dataset 1	gsnap	20	10	2	Total	3.29 ± 0.11	56.15 ± 0.96
dataset 1	rum	20	10	2	Rmsk	21.50 ± 0.17	65.55 ± 8.86
dataset 1	rum	20	10	2	Not in Rmsk	2.06 ± 0.05	64.39 ± 0.94
dataset 1	rum	20	10	2	Total	3.32 ± 0.08	64.52 ± 1.31
dataset 1	tophat2	20	10	2	Rmsk	18.90 ± 0.70	46.03 ± 4.65
dataset 1	tophat2	20	10	2	Not in Rmsk	1.70 ± 0.02	33.60 ± 1.83
dataset 1	tophat2	20	10	2	Total	2.61 ± 0.03	34.44 ± 1.48
dataset 1	gsnap	20	20	4	Rmsk	21.77 ± 0.67	50.99 ± 5.82
dataset 1	gsnap	20	20	4	Not in Rmsk	1.92 ± 0.04	60.59 ± 0.17
dataset 1	gsnap	20	20	4	Total	3.24 ± 0.11	59.96 ± 0.55
dataset 1	rum	20	20	4	Rmsk	21.33 ± 0.22	68.03 ± 8.08

Dataset	Aligner	Minimum Coverage	Minimum Level	Minimum RDD Count	Region	Percent of True Positives Removed	Percent of False Positives Removed
dataset 1	rum	20	20	4	Not in Rmsk	1.96 ± 0.05	70.78 ± 1.00
dataset 1	rum	20	20	4	Total	3.21 ± 0.08	70.64 ± 1.13
dataset 1	tophat2	20	20	4	Rmsk	18.56 ± 0.61	45.91 ± 2.00
dataset 1	tophat2	20	20	4	Not in Rmsk	1.65 ± 0.02	34.29 ± 1.45
dataset 1	tophat2	20	20	4	Total	2.54 ± 0.05	35.12 ± 1.16
dataset 2	gsnap	10	10	2	Rmsk	27.10 ± 0.48	30.25 ± 0.32
dataset 2	gsnap	10	10	2	Not in Rmsk	2.47 ± 0.06	15.38 ± 0.19
dataset 2	gsnap	10	10	2	Total	8.70 ± 0.05	20.32 ± 0.23
dataset 2	rum	10	10	2	Rmsk	25.82 ± 0.34	48.21 ± 0.12
dataset 2	rum	10	10	2	Not in Rmsk	2.62 ± 0.08	31.96 ± 0.16
dataset 2	rum	10	10	2	Total	8.31 ± 0.04	37.59 ± 0.11
dataset 2	tophat2	10	10	2	Rmsk	25.59 ± 0.47	27.53 ± 0.67
dataset 2	tophat2	10	10	2	Not in Rmsk	2.37 ± 0.08	4.18 ± 0.20
dataset 2	tophat2	10	10	2	Total	5.21 ± 0.07	7.60 ± 0.13
dataset 2	gsnap	10	20	2	Rmsk	27.09 ± 0.41	40.71 ± 1.57
dataset 2	gsnap	10	20	2	Not in Rmsk	2.42 ± 0.05	48.04 ± 0.86
dataset 2	gsnap	10	20	2	Total	8.67 ± 0.03	46.17 ± 0.34
dataset 2	rum	10	20	2	Rmsk	25.64 ± 0.25	60.79 ± 0.43
dataset 2	rum	10	20	2	Not in Rmsk	2.47 ± 0.06	58.33 ± 0.58
dataset 2	rum	10	20	2	Total	8.18 ± 0.03	59.05 ± 0.29
dataset 2	tophat2	10	20	2	Rmsk	25.37 ± 0.47	34.02 ± 1.71
dataset 2	tophat2	10	20	2	Not in Rmsk	2.29 ± 0.04	10.48 ± 0.30
dataset 2	tophat2	10	20	2	Total	5.12 ± 0.05	14.62 ± 0.46
dataset 2	gsnap	20	10	2	Rmsk	26.81 ± 0.70	37.56 ± 1.00
dataset 2	gsnap	20	10	2	Not in Rmsk	2.36 ± 0.08	31.41 ± 1.33
dataset 2	gsnap	20	10	2	Total	7.05 ± 0.08	33.02 ± 1.19
dataset 2	rum	20	10	2	Rmsk	25.35 ± 0.54	55.28 ± 0.74
dataset 2	rum	20	10	2	Not in Rmsk	2.52 ± 0.08	47.47 ± 0.11
dataset 2	rum	20	10	2	Total	6.73 ± 0.09	49.65 ± 0.14
dataset 2	tophat2	20	10	2	Rmsk	23.75 ± 0.98	32.07 ± 1.83
dataset 2	tophat2	20	10	2	Not in Rmsk	2.43 ± 0.09	7.71 ± 0.73
dataset 2	tophat2	20	10	2	Total	4.33 ± 0.16	10.58 ± 0.66
dataset 2	gsnap	20	20	4	Rmsk	26.81 ± 0.59	54.78 ± 0.95
dataset 2	gsnap	20	20	4	Not in Rmsk	2.31 ± 0.06	59.40 ± 2.00

Dataset	Aligner	Minimum Coverage	Minimum Level	Minimum RDD Count	Region	Percent of True Positives Removed	Percent of False Positives Removed
dataset 2	gsnap	20	20	4	Total	7.00 ± 0.06	58.54 ± 1.80
dataset 2	rum	20	20	4	Rmsk	25.13 ± 0.42	64.96 ± 0.89
dataset 2	rum	20	20	4	Not in Rmsk	2.37 ± 0.06	66.82 ± 0.34
dataset 2	rum	20	20	4	Total	6.56 ± 0.06	66.38 ± 0.13
dataset 2	tophat2	20	20	4	Rmsk	23.53 ± 1.07	49.60 ± 7.88
dataset 2	tophat2	20	20	4	Not in Rmsk	2.31 ± 0.05	23.43 ± 1.23
dataset 2	tophat2	20	20	4	Total	4.18 ± 0.13	27.49 ± 0.30

Table 4.16 Effect of BLAT filter, pseudogene filter, and removal of intronic sites within 6 bp of exon junctions on FDR of RDD detection.

Dataset	Aligner	Minimum Coverage	Minimum Level	Minimum RDD Count	Region	FDR Before BLAT Filter (%)	FDR After BLAT Filter (%)	Percent Decrease in FDR
dataset 1	gsnap	10	10	2	Rmsk	2.02 ± 0.18	1.20 ± 0.10	40.12 ± 6.15
dataset 1	gsnap	10	10	2	Not in Rmsk	1.92 ± 0.05	0.79 ± 0.03	59.03 ± 0.48
dataset 1	gsnap	10	10	2	Total	1.92 ± 0.06	0.81 ± 0.03	57.93 ± 0.75
dataset 1	rum	10	10	2	Rmsk	6.06 ± 0.71	2.70 ± 0.36	54.95 ± 8.50
dataset 1	rum	10	10	2	Not in Rmsk	5.41 ± 0.06	1.92 ± 0.02	64.55 ± 0.45
dataset 1	rum	10	10	2	Total	5.45 ± 0.10	1.96 ± 0.01	64.02 ± 0.59
dataset 1	tophat2	10	10	2	Rmsk	1.43 ± 0.17	0.96 ± 0.04	32.92 ± 5.86
dataset 1	tophat2	10	10	2	Not in Rmsk	1.24 ± 0.04	0.80 ± 0.03	35.76 ± 1.21
dataset 1	tophat2	10	10	2	Total	1.26 ± 0.04	0.81 ± 0.03	35.73 ± 1.38
dataset 1	gsnap	10	20	2	Rmsk	1.68 ± 0.20	0.96 ± 0.08	42.67 ± 4.34
dataset 1	gsnap	10	20	2	Not in Rmsk	1.61 ± 0.05	0.59 ± 0.03	63.40 ± 0.41
dataset 1	gsnap	10	20	2	Total	1.61 ± 0.06	0.61 ± 0.03	62.18 ± 0.37
dataset 1	rum	10	20	2	Rmsk	4.97 ± 0.73	2.05 ± 0.15	58.16 ± 6.81
dataset 1	rum	10	20	2	Not in Rmsk	4.47 ± 0.06	1.30 ± 0.01	70.91 ± 0.54
dataset 1	rum	10	20	2	Total	4.50 ± 0.09	1.34 ± 0.01	70.19 ± 0.48
dataset 1	tophat2	10	20	2	Rmsk	1.05 ± 0.13	0.72 ± 0.07	31.31 ± 2.05
dataset 1	tophat2	10	20	2	Not in Rmsk	0.87 ± 0.03	0.54 ± 0.02	37.86 ± 1.00
dataset 1	tophat2	10	20	2	Total	0.88 ± 0.03	0.55 ± 0.02	37.63 ± 1.04
dataset 1	gsnap	20	10	2	Rmsk	1.58 ± 0.12	1.01 ± 0.15	36.05 ± 7.82
dataset 1	gsnap	20	10	2	Not in Rmsk	1.55 ± 0.04	0.69 ± 0.03	55.32 ± 0.67
dataset 1	gsnap	20	10	2	Total	1.55 ± 0.05	0.71 ± 0.03	54.27 ± 1.05
dataset 1	rum	20	10	2	Rmsk	5.82 ± 0.84	2.58 ± 0.43	54.68 ± 11.28
dataset 1	rum	20	10	2	Not in Rmsk	4.95 ± 0.06	1.86 ± 0.03	62.46 ± 0.98
dataset 1	rum	20	10	2	Total	5.00 ± 0.11	1.89 ± 0.03	62.10 ± 1.38
dataset 1	tophat2	20	10	2	Rmsk	1.52 ± 0.21	1.01 ± 0.06	33.12 ± 5.52
dataset 1	tophat2	20	10	2	Not in Rmsk	1.21 ± 0.03	0.82 ± 0.04	32.18 ± 1.87
dataset 1	tophat2	20	10	2	Total	1.23 ± 0.04	0.83 ± 0.04	32.41 ± 1.53
dataset 1	gsnap	20	20	4	Rmsk	1.20 ± 0.10	0.76 ± 0.11	37.03 ± 7.98
dataset 1	gsnap	20	20	4	Not in Rmsk	1.23 ± 0.04	0.50 ± 0.02	59.52 ± 0.19
dataset 1	gsnap	20	20	4	Total	1.23 ± 0.05	0.51 ± 0.02	58.32 ± 0.62
dataset 1	rum	20	20	4	Rmsk	4.54 ± 0.86	1.84 ± 0.21	58.28 ± 10.25
dataset 1	rum	20	20	4	Not in Rmsk	3.90 ± 0.06	1.20 ± 0.02	69.35 ± 1.03
dataset 1	rum	20	20	4	Total	3.95 ± 0.09	1.23 ± 0.02	68.81 ± 1.19
dataset 1	tophat2	20	20	4	Rmsk	1.11 ± 0.13	0.74 ± 0.06	33.34 ± 2.08
dataset 1	tophat2	20	20	4	Not in Rmsk	0.82 ± 0.02	0.55 ± 0.03	33.00 ± 1.48

Dataset	Aligner	Minimum Coverage	Minimum Level	Minimum RDD Count	Region	FDR Before BLAT Filter (%)	FDR After BLAT Filter (%)	Percent Decrease in FDR
dataset 1	tophat2	20	20	4	Total	0.83 ± 0.03	0.56 ± 0.03	33.24 ± 1.22
dataset 2	gsnap	10	10	2	Rmsk	18.77 ± 0.17	18.10 ± 0.20	3.53 ± 0.17
dataset 2	gsnap	10	10	2	Not in Rmsk	13.57 ± 0.13	11.99 ± 0.14	11.65 ± 0.23
dataset 2	gsnap	10	10	2	Total	14.94 ± 0.10	13.30 ± 0.11	11.04 ± 0.23
dataset 2	rum	10	10	2	Rmsk	37.54 ± 0.27	29.56 ± 0.28	21.26 ± 0.28
dataset 2	rum	10	10	2	Not in Rmsk	26.95 ± 0.26	20.50 ± 0.26	23.96 ± 0.24
dataset 2	rum	10	10	2	Total	29.87 ± 0.24	22.48 ± 0.22	24.75 ± 0.16
dataset 2	tophat2	10	10	2	Rmsk	36.97 ± 0.07	36.36 ± 0.19	1.66 ± 0.59
dataset 2	tophat2	10	10	2	Not in Rmsk	32.34 ± 0.15	31.93 ± 0.17	1.26 ± 0.09
dataset 2	tophat2	10	10	2	Total	32.94 ± 0.13	32.38 ± 0.14	1.70 ± 0.07
dataset 2	gsnap	10	20	2	Rmsk	3.48 ± 0.05	2.84 ± 0.04	18.15 ± 2.19
dataset 2	gsnap	10	20	2	Not in Rmsk	3.45 ± 0.08	1.87 ± 0.07	45.88 ± 0.90
dataset 2	gsnap	10	20	2	Total	3.46 ± 0.07	2.07 ± 0.05	40.21 ± 0.38
dataset 2	rum	10	20	2	Rmsk	12.91 ± 0.33	7.25 ± 0.24	43.83 ± 0.76
dataset 2	rum	10	20	2	Not in Rmsk	10.48 ± 0.09	4.77 ± 0.10	54.55 ± 0.65
dataset 2	rum	10	20	2	Total	11.09 ± 0.13	5.27 ± 0.09	52.48 ± 0.35
dataset 2	tophat2	10	20	2	Rmsk	8.94 ± 0.09	7.99 ± 0.17	10.66 ± 2.46
dataset 2	tophat2	10	20	2	Not in Rmsk	6.04 ± 0.06	5.56 ± 0.05	7.91 ± 0.26
dataset 2	tophat2	10	20	2	Total	6.40 ± 0.06	5.80 ± 0.03	9.43 ± 0.47
dataset 2	gsnap	20	10	2	Rmsk	7.04 ± 0.18	6.06 ± 0.14	13.80 ± 0.82
dataset 2	gsnap	20	10	2	Not in Rmsk	4.83 ± 0.08	3.44 ± 0.08	28.73 ± 1.39
dataset 2	gsnap	20	10	2	Total	5.26 ± 0.09	3.85 ± 0.06	26.86 ± 1.20
dataset 2	rum	20	10	2	Rmsk	25.42 ± 0.49	16.96 ± 0.68	33.30 ± 1.36
dataset 2	rum	20	10	2	Not in Rmsk	16.60 ± 0.21	9.69 ± 0.15	41.64 ± 0.18
dataset 2	rum	20	10	2	Total	18.38 ± 0.21	10.84 ± 0.17	41.02 ± 0.25
dataset 2	tophat2	20	10	2	Rmsk	18.00 ± 0.74	16.37 ± 1.16	9.12 ± 3.08
dataset 2	tophat2	20	10	2	Not in Rmsk	13.79 ± 0.28	13.14 ± 0.35	4.70 ± 0.64
dataset 2	tophat2	20	10	2	Total	14.18 ± 0.30	13.38 ± 0.37	5.66 ± 0.60
dataset 2	gsnap	20	20	4	Rmsk	1.98 ± 0.16	1.23 ± 0.09	37.75 ± 0.87
dataset 2	gsnap	20	20	4	Not in Rmsk	2.02 ± 0.12	0.85 ± 0.05	57.95 ± 2.06
dataset 2	gsnap	20	20	4	Total	2.02 ± 0.12	0.91 ± 0.05	54.92 ± 1.92
dataset 2	rum	20	20	4	Rmsk	10.65 ± 0.54	5.29 ± 0.41	50.38 ± 1.57
dataset 2	rum	20	20	4	Not in Rmsk	8.18 ± 0.10	2.94 ± 0.07	64.08 ± 0.40
dataset 2	rum	20	20	4	Total	8.65 ± 0.13	3.29 ± 0.07	61.91 ± 0.20
dataset 2	tophat2	20	20	4	Rmsk	4.40 ± 0.30	2.93 ± 0.36	33.13 ± 9.91
dataset 2	tophat2	20	20	4	Not in Rmsk	2.39 ± 0.05	1.88 ± 0.04	21.21 ± 1.23
dataset 2	tophat2	20	20	4	Total	2.57 ± 0.06	1.96 ± 0.05	23.84 ± 0.32

4.3.7 Effect of non-random sequencing errors on FDR of RDD detection

Our analysis of the simulated RNA-Seq datasets allowed us to evaluate the relative contribution of alignment and sequencing error on RDD detection. In particular, we assumed a sequencing error profile in which errors are generated and independently. However, previous studies have indicated that the sequencing errors introduced by Illumina next-generation sequencing platforms may occur in a sequence-specific or non-independent manner (Minoche et al. 2011; Nakamura et al. 2011). To estimate the effect of such non-random and non-independent sequencing errors on the identification of RDDs, we analyzed two replicates of a deeply sequenced dataset (approximately 2,000 to 3,000x coverage on average) comprising 1,062 human cDNA clones (see Materials and Methods). We observed an overall error rate of 3.8×10^{-4} and 1.2×10^{-3} in the two datasets respectively across sites with a minimum coverage of 1,000x. Furthermore, we found that these errors are distributed in neither a random nor independent way (see Materials and Methods). In particular, we calculated that the frequency of errors introduced at levels of 20% or greater is 3.12×10^{-5} (average across two replicates). This frequency is comparable to the numbers of RDDs previously reported by others. As our methods for RDD detection cannot discriminate between such errors and RDDs, these non-random errors will play a nontrivial impact on the FDR of RDD detection.

4.3.8 Evaluation of RDDs in human lymphoblastoid cell line

Lastly, to evaluate the performance of our pipeline on a real experimental dataset, we analyzed the human lymphoblastoid cell line GM12878, for which deep DNA and RNA sequence is readily available (Dunham et al. 2012). We used the parameters and thresholds as determined from our previous synthetic data analyses to identify RDDs. In particular, we aligned two replicates of RNA-Seq data containing approximately 120 million 76 bp paired-end reads each using GSNAP (Table 4.17) and identified RDDs using a 'minimum coverage of 20x, minimum level of 20%, and minimum of 4 reads containing the sequence difference base' threshold. Sequence differences found in dbSNP137 (Sherry et al. 2001) were removed from consideration. Furthermore, to minimize the detection of sequence differences resulting from sequencing error, we focused our analysis on those differences that are observed in both replicates (Table 4.18). Overall, we preliminarily identified a total of 12,480 RDDs in the two replicates. We separated the differences by type into two groups: A-to-G sequence differences and noncanonical sequence differences, or changes that cannot be explained by known mechanisms. We note that although C-to-T differences can be mediated by APOBEC, APOBEC1 is not expressed in this B-cell cell line, with an FPKM value (Trapnell et al. 2010) of 0 in both replicates. The majority (56.38%) of the sequence differences we identified are A-to-G changes and likely to be mediated by RNA editing via ADAR.

Table 4.17 Alignment statistics for GM12878 RNA-Seq dataset.

Replicate	Statistic	Value
1	Number of Read Pairs Sequenced	117,876,320
1	Number of Read Pairs Aligned	106,890,066
1	Percentage of Read Pairs Aligned	90.68%
1	Number of Read Pairs Aligned Uniquely	100,040,977
1	Percentage of Read Pairs Aligned Uniquely	93.59%
2	Number of Read Pairs Sequenced	131,831,897
2	Number of Read Pairs Aligned	124,158,569
2	Percentage of Read Pairs Aligned	94.18%
2	Number of Read Pairs Aligned Uniquely	115,844,755
2	Percentage of Read Pairs Aligned Uniquely	93.30%

Table 4.18 RNA-DNA sequence differences found in GM12878.

	Both	Replicate 1	Replicate 2	Total
A>C	646	2,536	1,580	4,762
A>G	7,036	5,739	10,775	23,550
A>T	131	354	128	613
C>A	166	360	166	692
C>G	286	654	505	1,445
C>T	492	843	414	1,749
G>A	563	856	321	1,740
G>C	200	546	330	1,076
G>T	169	488	197	854
T>A	132	362	137	631
T>C	1,300	2,713	1,843	5,856
T>G	1,359	3,488	2,636	7,483
Total	12,480	18,939	19,032	50,451

We investigated whether filters we previously identified as effective in removing false positive RDDs could explain the sequence differences we observed. Other researchers have used these filters in their pipelines to accurately identify RDDs (Kleinman et al. 2012; Peng et al. 2012; Ramaswami et al. 2012). The filters we implemented include searching with BLAT for regions homologous to the sequence flanking the sequence difference (see Materials and Methods), removing intronic sites near exon-exon junctions, and eliminating differences in annotated pseudogenes or adjacent to homopolymer sequences. We applied the BLAT filter to sequence differences found outside of RepeatMasker regions, as we previously showed that this filter is not as effective in discriminating between true and false positives within repetitive sequences. We observed that approximately 47% of noncanonical differences are removed by one or more of these filters, whereas only roughly 14% of A-to-G sites are eliminated (Table 4.19). The filtering steps that filtered out the greatest percentage of sites are the pseudogene and BLAT filters, as nearly 30% of noncanonical sites are removed by each filter independently. After taking into consideration all of the filters we used, a total of 8,933 sequence differences remained, 68% of which are A-to-G edits (Figure 4.11). Of these 8,933 differences, the majority (72%) are located within RepeatMasker regions. Within RepeatMasker regions, nearly 87% of the differences are A-to-G, as is expected due to the phenomenon of editing in human *Alu* elements (Athanasiadis et al. 2004). In contrast, the majority (81%) of sites outside of RepeatMasker are noncanonical differences. The distribution of sequence differences we observed are highly concordant with other studies, and the most common noncanonical RDD types we observed were A-

to-C and its complement T-to-G, as previously seen by others (Ju et al. 2011; Kleinman et al. 2012; Ramaswami et al. 2012).

For the remaining RDDs that are not removed by our filtering methods, we asked whether features indicative of sequencing error or low-quality mapping are more common in noncanonical versus A-to-G sequence differences. Specifically, we noticed that many noncanonical sequence differences occur within regions where many of the reads overlapping the sequence difference site are either partially mapped via a local alignment with clipped bases or mapped with many mismatches (Figure 4.12). To investigate the mapping quality near sites of RDDs globally, we calculated for each read that overlaps an RDD the number of bases (out of the total 76 bp sequence) that are neither clipped nor aligned with a mismatch or indel; we refer to this figure as the number of bases aligned properly. We observed that in both replicates, for sequence differences in RepeatMasker, the overall number of bases that are aligned properly is higher for A-to-G changes than for the most of the noncanonical types (Figure 4.13). For sites lying outside of RepeatMasker regions, we observe that the number of bases aligned properly for noncanonical sequence differences, excluding the more common A-to-C and T-to-G types, is generally lower than for that of A-to-G (Figure 4.12).

Table 4.19 Number of RDDs removed by various bioinformatics filters.

	A>G	Non-canonical	Total
Total before filters	**7,036**	**5,444**	**12,480**
Pseudogene filter (removed)	847 (12.04%)	1,959 (35.98%)	2,806
BLAT filter (removed)	545 (7.75%)	1,722 (31.63%)	2,267
Homopolymer filter (removed)	32 (0.45%)	205 (3.77%)	237
Exon junction filter (removed)	30 (0.43%)	88 (1.62%)	118
Total after filters (remaining)	**6,039 (85.83%)**	**2,894 (53.16%)**	**8,933**
Total after filters - in RmskRM327	5,560	837	6,397
Total after filters - not in RmskRM327	479	2,057	2,536

Figure 4.11 Distribution of RNA-DNA sequence differences in GM12878. Here we depict the distribution of RDDs in GM12878 after removing sites using various filters.

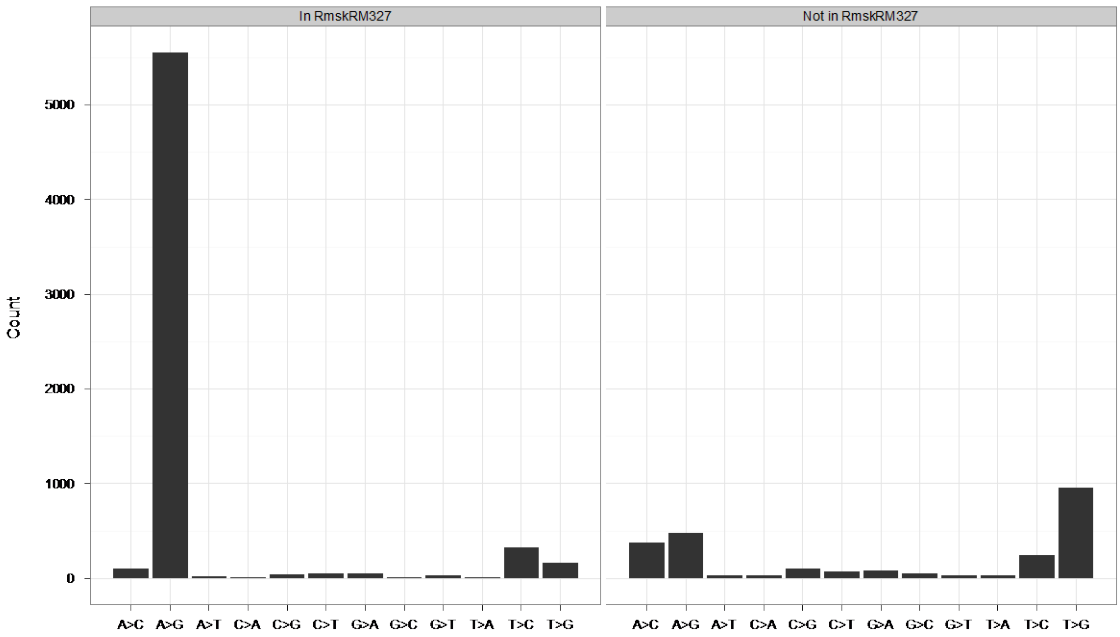


Figure 4.12 T-to-G RNA-DNA sequence difference at chr10:102046378 (hg19). Here we show an image in the IGV browser (Thorvaldsdottir et al. 2012) of a T-to-G sequence differences at chr10:102046378 in the first replicate of the GM12878 dataset. Each grey bar represents an RNA-Seq read. Mismatches are depicted by colored letters. Black dashes within a read represent a clipped sequence; for reads in the bottom half, the string of colored bases depict clipped portions of the sequence. Clipped portions of alignments represent bases that are not aligned within a local alignment.

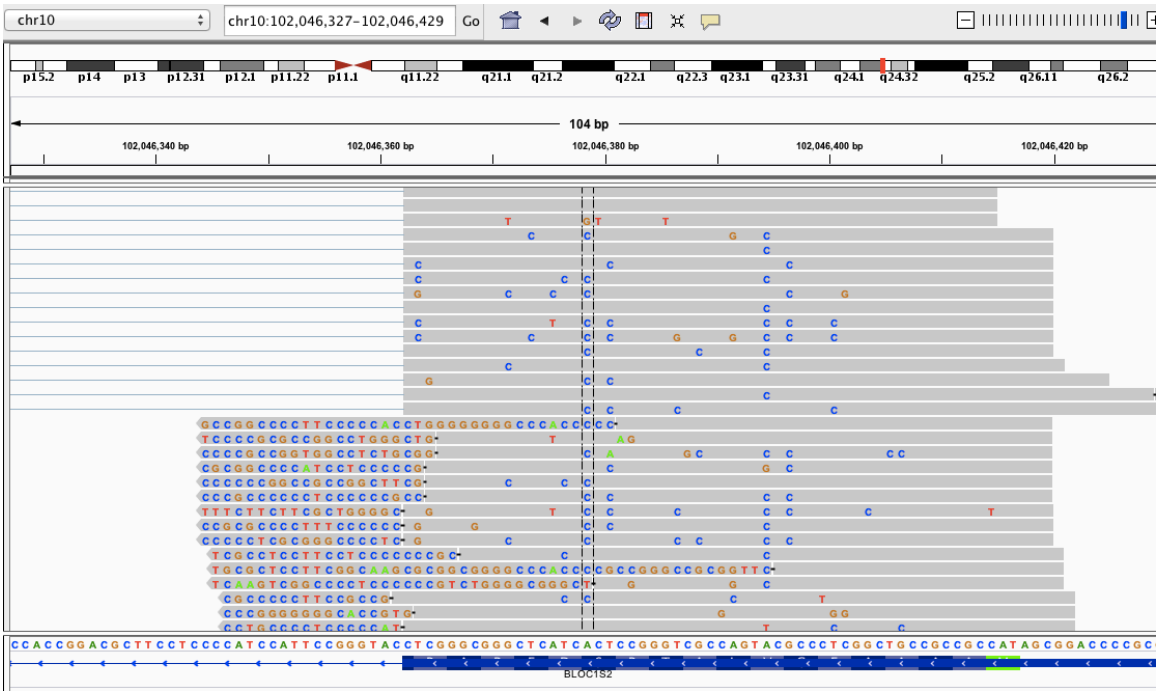
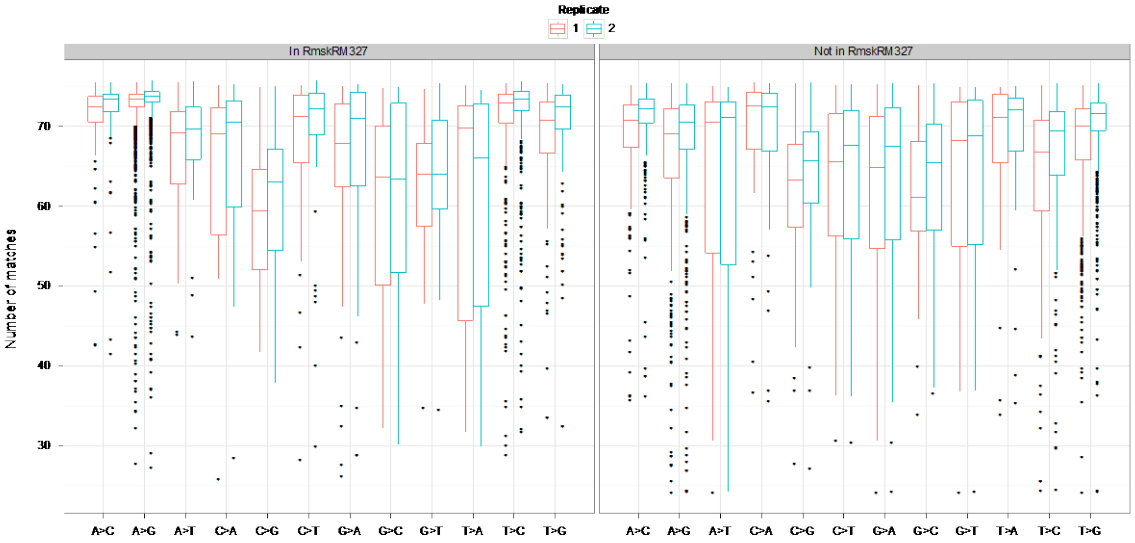


Figure 4.13 Number of properly aligned bases in reads that overlap RDDs. Here we depict the number of bases within each read that overlaps an RNA-DNA sequence difference site that are aligned properly. This number excludes bases that contain mismatches or those that are clipped or part of an insertion or deletion.



4.4 Discussion

RNA-Sequencing is a powerful technology for genome-wide analyses of transcriptome information at the single-nucleotide level. The resolution afforded by next-generation sequencing technology has allowed for genome-wide studies on RNA editing in humans (J. B. Li et al. 2009; Peng et al. 2012) and led to the identification of all 12 types of sequence differences (M. Li et al. 2011). There are, however, limitations to high-throughput sequencing, as difficulties lie in the alignment of short sequencing reads and errors introduced by sequencing and library preparation among other challenges. The relative effect of these various misalignment and sequencing errors on the identification of RDDs is debated, although many reports assert that the majority of noncanonical sequence differences observed result from technical artifacts (Kleinman & Majewski 2012; W. Lin et al. 2012; Pickrell et al. 2012; Schrider et al. 2011). In this study, we dissect the various sources of error leading to false positive RDDs and evaluate their relative contribution. Using a detection theory approach, we generated simulated RNA-Seq datasets containing known RDDs to evaluate the effect of alignment and sequencing error on RDD analysis. In the absence of sequencing error, we found that minimal thresholds are sufficient for sensitivity values above 95% and false discovery rates below 5%. Moreover, we found that the RDD levels reported by the various aligners correlate well ($R \sim 98\%$) with the true levels per our simulation. Upon introduction of sequencing errors following a random and independent distribution, we found that a threshold requiring a ‘minimum coverage of 20x, minimum level of 20%, and minimum of 4 reads bearing the RDD base’ is necessary for false discovery rates below 10% across the various aligners. In addition to random and independent sequencing errors, we evaluated

the effect of non-random sequencing errors on the analysis of RDDs and found that they play a large part in the inflation of false discovery rates.

Currently, most pipelines use ad hoc filtering methods to minimize the presence of false positives in sequence difference studies without a full understanding of the efficacy of these methods or the trade-off between sensitivity and false discovery rates. We found that overall while the various filters used in the literature for removal of false positive RDDs are effective in discriminating between true and false positives, a sizeable percentage of false positives remain even after all filtering methods are implemented.

Lastly, we used our pipeline for identification of RDDs to evaluate the presence of sequence differences in humans. Using parameters and thresholds we deemed as optimal, we identified approximately 9,000 RDDs, the majority (68%) of which are A-to-G changes and likely to be mediated by ADAR. Of the noncanonical RDDs that remained after our filtering processes, we found A-to-C and its complement T-to-G to be most common. Notably, A-to-C changes have been found by others to be the most common sequencing error (Dohm et al. 2008; Qu et al. 2009). Furthermore, we found that the alignments of reads overlapping noncanonical RDDs, with the exception of A-to-C and T-to-G types, contain many more mismatches or clipped bases than those of A-to-G differences. The distribution of sequence differences we observed is highly concordant with previous studies, and like others (Kleinman et al. 2012; Ramaswami et al. 2013), we conclude that there is little evidence for widespread noncanonical editing.

Overall, we observed that next-generation sequencing technology and current bioinformatics tools are a reliable and powerful technique for studying RDDs genome-wide. Furthermore, we found that computational biology methods are an effective means

for evaluating the various thresholds and filtering techniques used to accurately identify sequence differences. Our results demonstrate that while RNA-Seq allows for precise detection and measurement of RDDs, current bioinformatics filters do not completely remove false positive calls. We aim for this study to provide a general framework for those interested in site-specific allelic differences in humans using RNA-Seq, and hope in particular that our work may shed light on the appropriate thresholds and necessary caution to employ for RDD analyses.

4.5 Materials and Methods

4.5.1 Simulation of RNA-Seq datasets

Simulated datasets were generated using the BEERS simulator (Grant et al. 2011). Data are based on human build hg19 and RefSeq transcript models (Pruitt et al. 2012), as aligned to the genome by UCSC (Kent et al. 2002) using BLAT (Kent 2002). The expression intensities are Poisson distributed with probabilities estimated from roughly 300 million reads of human retina RNA-Seq data, as described previously (Grant et al. 2011). Default settings result in 36,467 transcripts, of which approximately 70% are expressed. We simulated two types of RNA-Seq datasets. Dataset 1 was "clean" and designed to contain no intron signal or sequencing error. Dataset 2 was "realistic" and constructed with intron retention and sequencing error. We used a substitutional error rate of 1 in 200 (0.5%), a value comparable to observed sequencing error rates of Illumina Genome Analyzer IIx and HiSeq machines (Minoche et al. 2011). Furthermore, we simulated poorer quality bases at the ends of reads by increasing the substitutional error rate to 20% in the last 10 bases for 25% of the reads. Approximately 30% of the signal in the dataset originates from introns. These parameters are consistent with real data observations. Lastly, we also included indel polymorphisms at a rate of 1 in 1000 (0.1%). Both datasets 1 and 2 were generated in triplicate, with each replicate containing 50 million pairs of reads of length 100 base pairs (bp). The mean fragment length of each read pair is 330 bp.

4.5.2 Alignment of RNA-Seq datasets

RNA-Seq datasets were aligned using GSNAP version 2012-07-20 (T. D. Wu & Nacu 2010), RUM version 2.0.3-02 (Grant et al. 2011), or Tophat2 version 2.0.6 (Trapnell et al. 2009) to the human genome (build hg19). GSNAP was run with default options. A maximum number of 10 alignments were permitted for each read. Alignments to novel exon-exon junctions (per GSNAP option -N 1) and known junctions as defined by RefSeq (downloaded November 2, 2012) and Gencode version 13 (Harrow et al. 2006) were accepted. Alignments with no more than the default maximum of '(read length + 2)/12 – 2' mismatches were retained. RUM was run with the default command line options. Tophat2 was run with the default options. A maximum edit distance and mismatch count of 6 was allowed for each read. Secondary alignments up to the default maximum of 20 were permitted. After alignment with GSNAP, RUM, or Tophat2, non-primary alignments and alignments placing read pairs in the incorrect orientation were removed.

4.5.3 Simulation of RNA-DNA sequence differences

For each dataset, sites in the genome are first stratified by coverage to ensure the placement of RDDs at locations with varying depths of coverage. The distribution of coverage for dataset 1, which does not contain reads originating from intronic regions of the genome, is fairly uniform, while for dataset 2, the distribution is skewed right (Figure S2); approximately 82% of sites in dataset 2 have coverage of 10x or less compared to approximately 20% in dataset 1. For dataset 1, we grouped sites into quartiles, corresponding to coverage values of approximately 0x to 14x for quartile 1, 15x to 49x for quartile 2, 50x to 133x for quartile 3, and 134x and above for quartile 4. For dataset 2,

the presence of introns results in a highly skewed right distribution for coverage. As such, we divided dataset 2 into one group containing sites with coverage below 10x and split the remaining sites into tertiles, corresponding to coverage values of approximately 11x to 19x for tertile 1, 20x to 48x for tertile 2, and 49x and above for tertile 3. After we grouped sites by coverage, we randomly inserted RDDs at different sites such that each coverage group contained approximately the same number of sequence differences.

The type of RDD difference (e.g. A-to-C, A-to-G, A-to-T, etc.) was determined randomly and independently for each site. The RDD level, or the proportion of reads containing the sequence difference, was chosen randomly from a random uniform distribution from 0 to 1, excluding 0.

A small subset (5%) of the simulated RDDs was randomly chosen to model hyperediting, or the clustering of many sequence differences in a small window. In particular, we designated all of the sites that are within 100 bp of the chosen site to have a 50% chance of having the same RDD type provided that the coverage belongs to the same coverage group as the initial site.

4.5.4 Repetitive regions of the genome as defined by BLAT

As one measure of the repetitive nature of a region surrounding a sequence difference site, we used BLAT (Kent 2002) to search for homologous sequences in the genome. In particular, we extracted flanking sequences of length 51 bp, 101 bp, and 151 bp around a given site and queried for alignments in the genome with BLAT (v.35x1). The settings `--stepSize=5` and `--repMatch=2253` were used to increase sensitivity. A maximum of `(read length + 2)/12 – 2` mismatches per alignment, the same amount

permitted by GSNAP, was tolerated. Sites for which more than one alignment is found for one of the three flanking sequences are deemed “non-unique by BLAT”.

4.5.5 Filtering of RNA-DNA sequence differences using BLAT

To ensure that an RDD identified by the various aligners cannot be explained by homologous sequences in the genome, sequences of length 25, 50, and 75 bp upstream and downstream of each sequence difference site were aligned to the genome using BLAT (v. 35x1). The settings `--stepSize=5` and `--repMatch=2253` were used to increase sensitivity. A maximum of `(read length + 2)/12 – 2` mismatches per alignment, the same amount allowed by GSNAP, was tolerated. An RDD was filtered out if any of the flanking sequences aligned to a region other than the RDD site and if that alignment explained the sequence difference.

4.5.6 Analysis of non-random sequencing errors in experimental RNA-Seq datasets

Note: Experimental work in this section was performed by Nicholas Lahens in the John Hogenesch laboratory.

To evaluate the extent to which non-random sequencing errors can affect the false discovery rate of RDD detection, we analyzed two replicates of a dataset comprising 1,062 cDNAs from the Mammalian Genome Collection (Temple et al. 2009) that were expressed *in vitro* and sequenced using Illumina HiSeq 2000 technology. RNA was treated with Ribo-Zero Gold kit (Epicentre catalog no. RZHM11106) and converted into an Illumina RNA-Seq library with the TruSeq RNA sample prep kit (Ilumina catalog no. FC-122-1001). Briefly, rRNA was removed from 1 ug of IVT RNA using Ribo-Zero Gold kit and purified via ethanol/sodium acetate precipitation according to manufacturers

protocol. After drying, the RNA pellet was dissolved in 18 μL of Elute, Prime, Fragment mix (provided with TruSeq RNA sample prep kit). RNA was fragmented for 8 minutes and 17 uL of this fragmented RNA was used to make the RNA-Seq library according to Illumina RNA TruSeq RNA sample prep kit protocol. After fragmentation/priming, first strand cDNA synthesis with SuperScript II (Invitrogen catalog no. 18064014), second-strand synthesis, end-repair, a-tailing, and adapter ligation, the library fragments were enriched with 15 cycles of PCR. Quality and size of library was assessed using Agilent 2100 BioAnalyzer. Library was sequenced in replicate using Illumina HiSeq 2000 to obtain approximately 41 million and 32 million 100-bp paired-end reads.

We aligned both replicates using GSNAP (see 'Alignment of RNA-Seq datasets' in Materials and Methods) to an index containing the non-spliced reference sequence of the 1,062 cDNAs. For each replicate, approximately 82% of the total reads were aligned in the correct orientation and with the expected inner distance between read pairs. The coverage distribution in each replicate is fairly uniform, with an average of approximately 2,600x and 3,000x in replicates 1 and 2, respectively (median of 2,300x in replicate 1 and 1,800x in replicate 2). To be confident of the sequencing and alignment results, we restricted our analyses to sites with a minimum coverage of 1,000x. In total, we obtained 4,923,509,994 and 4,195,153,516 bases of sequence at 1,209,658 and 1,111,552 sites and observed a total of 1,877,330 and 4,902,349 sequencing errors, giving an overall error rate of approximately 3.8×10^{-4} and 1.2×10^{-3} in replicates 1 and 2, respectively. For each site, we calculated the sequencing error level to be the percentage of total reads at the site bearing an error. To test whether the observed sequencing errors occur randomly, we performed a Kolmogorov-Smirnov test, comparing the observed

distribution of sequencing error levels to a null distribution derived from the overall error rate calculated previously. We found that for both replicates, the distribution of sequencing error levels deviates from that expected under the null distribution ($P < 0.001$ for both replicates).

Chapter 5. Genetic Basis of RNA-DNA Sequence Differences

5.1 Abstract

RNA-Sequencing (RNA-Seq) delivers quantitative and comprehensive information on transcriptomes at the single-nucleotide level. Such detailed information on DNA and RNA allows for systematic identification of RNA-DNA sequence differences (RDDs) genome-wide. Recent studies have utilized next-generation sequencing technology to survey the landscape of known RNA editing processes in humans as well as detect sequence differences between DNA and RNA that cannot be explained by known mechanisms. In this chapter, we explore the genetic basis of RDDs using RNA-Seq data. Little is known about whether RDD levels, or the proportion of reads at an RDD site that bear the sequence difference allele, vary among individuals genome-wide. Previous studies on ADAR editing in humans have observed consistent levels of editing across different samples in a few genes. In this study, we develop statistical methods for assessing the extent to which individuals vary in their RDD levels genome-wide using RNA-Seq data. We also quantify the degree to which such individual variation is genetically determined. In designing these statistical algorithms, we take into account variation in sequencing coverage that is present in next-generation sequencing data. We also correct for multiple-testing error as we envision our tests to be used on genome-sized datasets. We apply our statistical algorithms to RNA-Seq data for cultured human B-cells derived from 27 unrelated individuals and 10 pairs of monozygotic twins in the Centre d'Etude du Polymorphisme Humain (CEPH) collection. Among the 27 unrelated individuals, which were sequenced to an average coverage of 8x each, we observed over 120 sites in 60 genes for which A-to-G RNA editing levels vary across the unrelated

samples (FDR ~ 0.05). For the 10 pairs of monozygotic twins, which were sequenced to a depth of approximately 29x each, we found over 2,000 sites in roughly 500 genes for which variation in RDD levels is significantly explained by genetic differences. Our results provide the first genome-wide survey of variation and heritability of RDD levels in humans and our methods can be applied to future genetic analyses on RDDs and other quantitative phenotypes measured using RNA-Seq data.

5.2 Introduction

Next-generation sequencing technology provides comprehensive coverage of genomes and transcriptomes, facilitating comprehensive comparisons of DNA and RNA sequence. The transmission of sequence information from DNA to RNA is a critical process and is expected to occur in a one-to-one fashion; however, there are known exceptions. In humans, these exceptions encompass two types of RNA editing: adenosine to inosine edits as catalyzed by ADAR (Bass & Weintraub 1988) and cytosine to uridine changes as mediated by APOBEC1 (Powell et al. 1987). Inosine is recognized by the translational machinery as guanosine, and thus ADAR-mediated editing can result in changes in protein sequence. RNA editing by ADAR is essential for development and normal function in both invertebrates and vertebrates (Higuchi et al. 2000; Palladino et al. 2000). In addition, abnormalities in ADAR editing have been shown to be associated with various human disorders (Brusa et al. 1995; Kawahara et al. 2004; Silberberg et al. 2012).

Little is known about variation in levels of editing, the percentage of transcripts at an editing site that are edited, on a genome-wide scale. Previous reports have shown editing levels in glutamate receptors to be consistent across individuals (Paschen et al. 1994; Wahlstedt et al. 2009). Another study examined the levels of editing in *Alu* repeats for 6 different genes across 32 human skin samples and likewise found highly consistent levels of editing across different individuals and tissues (Greenberger et al. 2010). In this chapter, we propose methods that allow for genome-wide surveys of individual variation in RNA editing or RDD levels and, in addition, the degree to which such variation is explained by genetic determinants. The methods we develop are suited for the

quantitative nature of RNA-Seq data and account for variation in sequencing depths across different samples. Also, we design our algorithms with the goal of genome-wide analyses and thus correct for multiple testing error.

Our work aims to lay the foundation for future studies on the genetics of RDDs – both those for which the underlying mechanisms are known and those for which they are not. For RDDs generated by known enzymes such as ADAR and APOBEC1 editing, genetic analyses may shed light on the means by which such editing is regulated and controlled. Furthermore, with respect to noncanonical RDDs, a genetic approach represents one method for elucidation of the factors underlying the unknown mechanism. Lastly, differences in RDD levels across samples may also have functional consequences on cellular processes and may thus be an underlying determinant for individual phenotypic variation. With these motivations in mind, we propose methods for evaluating individual variation and heritability of RDD levels using next-generation sequencing data.

5.3 Results

5.3.1 Individual variation in RDD levels among unrelated individuals

We surveyed the landscape of ADAR RNA editing sites genome-wide to determine whether A-to-G editing or RDD levels vary among individuals. In particular, we obtained RNA-Seq data on 27 unrelated individuals from the Centre d'Etude du Polymorphisme Humain (CEPH) collection (Dausset et al. 1990). These are the same individuals previously used by Cheung and colleagues to identify RDDs in human B-cells (M. Li et al. 2011). For each individual, we aligned the RNA-Seq data using GSNAP and calculated RDD levels for A-to-G sites in the Database of RNA Editing in Humans (DARNED) (Kiran & Baranov 2010), a public collection of editing sites in humans (see Materials and Methods). To be confident of the RDD levels as measured by RNA-Seq data, we focused on sites where a sufficient number of individuals had relatively high coverage. In particular, we required a minimum of 10 individuals with coverage of at least 30x at each site under consideration. Overall, a total of 1,112 sites satisfied these criteria. We then asked whether the RDD levels at these sites varied across the 27 unrelated individuals (see Materials and Methods). Using a false positive rate of 0.005 (Figure 5.1), we identified a total of 120 sites (11%) in over 60 genes that showed significant individual variation in RDD levels (FDR ~ 0.05). A list of top sites demonstrating individual variation is shown in Table 5.1 along with a few examples in Figure 5.2.

Figure 5.1 False discovery rate versus false positive rate for identification of sites with significant individual variation in RDD levels. Here we calculate the false discovery rate versus false positive rate after correcting for multiple testing error in the test of individual variation in RDD levels across individuals (see Materials and Methods).

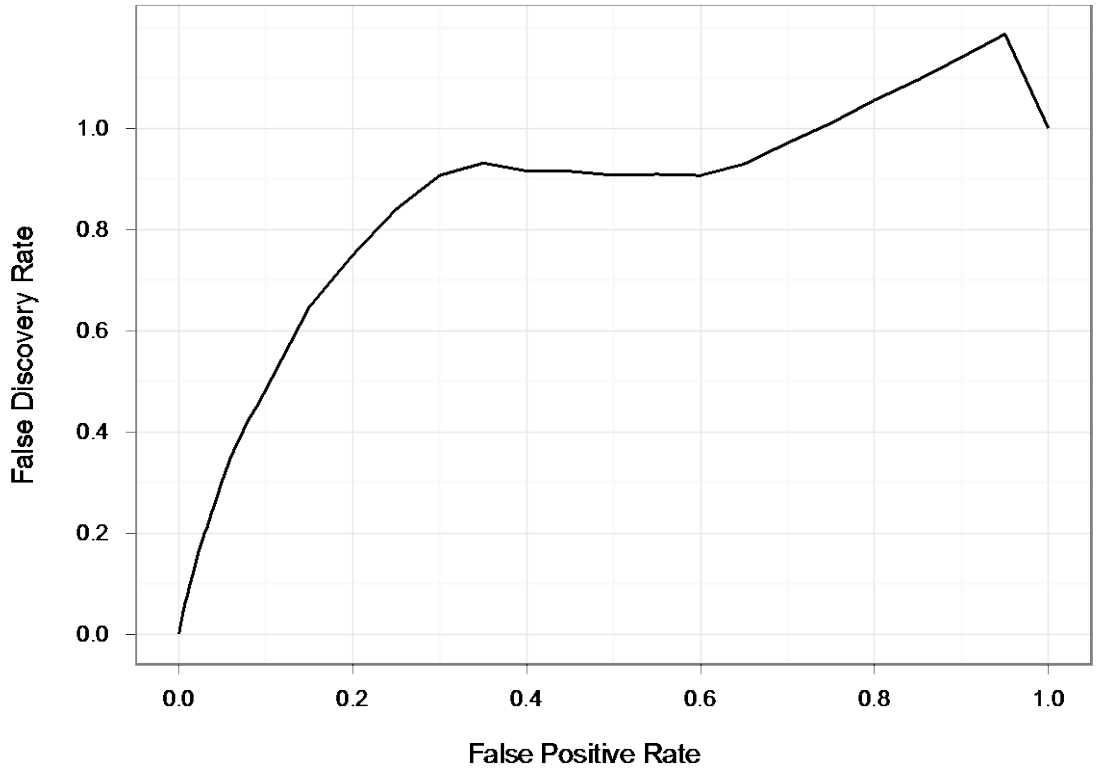


Table 5.1 List of top A-to-G RNA editing sites with significant variation in editing levels among 27 unrelated individuals.

chrom	position	gene name	feature	n§	mean RDD level	median RDD level	minimum RDD level	maximum RDD level	Z‡	p value
chr1	567242	intergenic	intergenic	27	1.36	1.16	0.28	4.17	111.73	0
chr1	1247494	CPSF3L	3UTR;CDS	22	70.75	100.00	0.00	100.00	871.71	0
chr1	6158562	KCNAB2	CDS	22	46.52	45.73	0.00	100.00	461.98	0
chr1	6159032	KCNAB2	3UTR	11	47.18	41.03	0.00	100.00	278.47	0
chr1	6160563	KCNAB2	3UTR	19	75.28	98.44	0.00	100.00	639.80	0
chr1	6160876	KCNAB2	3UTR	26	76.14	100.00	0.93	100.00	1613.90	0
chr1	6160958	KCNAB2	3UTR	27	70.58	100.00	0.00	100.00	1623.68	0
chr1	20978410	DDOST	3UTR	26	7.68	8.25	0.00	95.31	544.03	0
chr1	27210721	GPN2	CDS	13	68.32	54.84	0.00	100.00	273.86	0
chr1	31205796	LAPTM5	3UTR	27	33.45	40.41	0.00	100.00	16002.13	0
chr1	36067677	PSMB2	3UTR	21	21.91	24.69	3.90	46.51	90.68	0
chr1	38449910	SF3A3	CDS	23	21.95	0.00	0.00	100.00	747.22	0
chr1	46078854	NASP	3UTR;CDS	16	18.90	20.05	0.00	100.00	282.77	0
chr1	79108193	IFI44L	3UTR	25	1.00	0.00	0.00	9.68	146.76	0
chr1	79108193	IFI44L	3UTR	25	1.00	0.00	0.00	9.68	146.76	0
chr1	146650916	FMO5;PDIA3P	exon	27	64.56	42.68	0.00	100.00	2622.82	0
chr1	154897350	PMVK	3UTR	27	1.47	0.89	0.00	11.58	76.38	0
chr1	160302244	COPA	CDS	20	1.13	0.00	0.00	15.38	46.92	0
chr1	160967938	F11R	3UTR	13	22.88	20.37	0.00	39.39	35.72	0.001
chr1	160967938	F11R	3UTR	13	22.88	20.37	0.00	39.39	35.72	0.001
chr1	184761188	FAM129A	3UTR	14	11.01	8.96	0.00	27.66	45.32	0
chr1	220427469	MORF4L1P1	exon	27	10.58	0.00	0.00	53.42	1189.44	0
chr11	35827958	TRIM44	3UTR	21	13.32	0.00	0.00	100.00	611.13	0
chr11	60609972	CCDC86	CDS	13	12.90	0.00	0.00	52.78	230.96	0
chr11	77790653	NDUFC2	CDS	16	74.84	98.06	0.00	100.00	483.68	0
chr12	32880193	DNM1L	exon	27	44.12	54.21	0.00	100.00	5722.95	0
chr12	69237043	CPM;MDM2	3UTR	10	13.18	15.53	0.00	26.67	32.63	0.001
chr12	69237043	CPM;MDM2	3UTR	10	13.18	15.53	0.00	26.67	32.63	0
chr14	20916958	OSGEP	5UTR;CDS	18	33.80	35.49	0.00	100.00	717.23	0
chr14	93407583	ITPK1	3UTR	12	4.07	0.00	0.00	50.00	117.38	0
chr16	29679362	SPN	3UTR	27	6.39	5.41	0.00	15.05	76.81	0
chr16	29680796	SPN	3UTR	24	62.24	61.42	36.11	73.91	51.65	0.001

chrom	position	gene name	feature	n§	mean RDD level	median RDD level	minimum RDD level	maximum RDD level	Z‡	p value
chr16	29680922	SPN	3UTR	23	12.68	12.90	0.00	26.83	51.17	0.003
chr16	29681216	SPN	3UTR	25	21.09	21.43	5.68	30.91	50.10	0.003
chr16	29681216	SPN	3UTR	25	21.09	21.43	5.68	30.91	50.10	0.004
chr16	29681303	SPN	3UTR	22	20.22	19.29	0.00	40.00	49.07	0
chr16	29681303	SPN	3UTR	22	20.22	19.29	0.00	40.00	49.07	0.002
chr16	57399547	CCL22	3UTR	27	0.20	0.08	0.00	1.74	53.24	0.001
chr17	1373518	MYO1C	CDS	14	62.45	49.77	31.43	100.00	304.16	0
chr17	37920169	IKZF3	3UTR	14	40.72	42.21	15.69	65.00	36.11	0.001
chr17	37920635	IKZF3	3UTR	19	40.61	43.59	18.18	55.26	45.16	0
chr17	37920635	IKZF3	3UTR	19	40.61	43.59	18.18	55.26	45.16	0
chr17	80445942	NARF	exon	16	3.05	0.00	0.00	15.38	52.94	0
chr19	2072038	MOB3A	3UTR	26	36.35	43.18	0.00	100.00	2045.37	0
chr19	4362691	SH3GL1	CDS	19	11.40	0.00	0.00	55.00	344.37	0
chr19	14721449	CLEC17A	3UTR	25	37.58	35.29	18.52	56.96	90.12	0
chr19	14721449	CLEC17A	3UTR	25	37.58	35.29	18.52	56.96	90.12	0
chr19	14721471	CLEC17A	3UTR	25	4.19	2.99	0.00	16.18	54.49	0.002
chr19	14721471	CLEC17A	3UTR	25	4.19	2.99	0.00	16.18	54.49	0.003
chr19	18288551	IFI30	CDS	27	1.02	0.61	0.00	3.50	142.67	0
chr19	39322087	ECH1	CDS	26	44.86	40.60	0.00	100.00	1933.75	0
chr19	39359229	RINL	3UTR	23	15.78	16.33	3.33	37.21	51.52	0.001
chr19	39359229	RINL	3UTR	23	15.78	16.33	3.33	37.21	51.52	0.001
chr19	48255804	GLTSCR2	CDS	26	42.64	38.08	0.00	100.00	2179.54	0
chr19	48258717	GLTSCR2	CDS	26	60.37	38.65	0.00	100.00	2549.66	0
chr19	49470064	FTL	3UTR	27	0.37	0.23	0.00	1.48	110.98	0
chr2	87423754	ANAPC1	exon	17	11.48	0.00	0.00	53.33	287.40	0
chr2	127808046	BIN1	CDS	13	2.63	0.00	0.00	35.85	106.55	0
chr20	44054349	PIGT	CDS	10	45.86	44.12	0.00	100.00	388.38	0
chr21	34923319	SON	CDS	22	5.24	3.93	0.00	15.63	57.23	0.001
chr22	39414754	APOBEC3C	3UTR	27	3.53	2.22	0.00	10.73	153.65	0
chr22	39414754	APOBEC3C	3UTR	27	3.53	2.22	0.00	10.73	153.65	0
chr22	39414772	APOBEC3C	3UTR	27	14.26	12.09	0.91	38.78	395.52	0
chr22	39414772	APOBEC3C	3UTR	27	14.26	12.09	0.91	38.78	395.52	0
chr3	31678420	STT3B	3UTR	24	2.22	0.00	0.00	37.50	165.23	0
chr3	124802231	SLC12A8	3UTR	24	50.63	50.84	0.00	100.00	2068.93	0

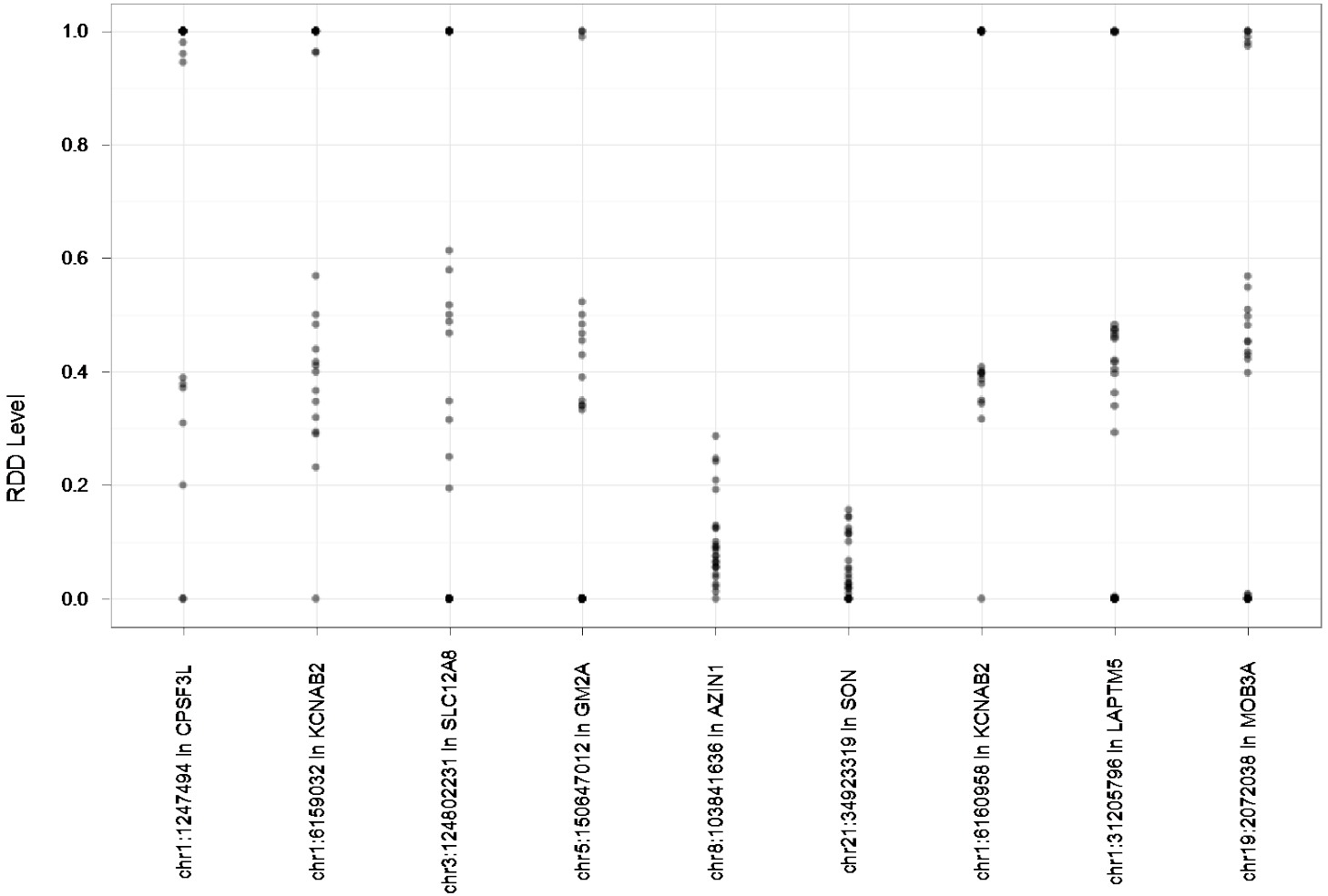
chrom	position	gene name	feature	n§	mean RDD level	median RDD level	minimum RDD level	maximum RDD level	Z‡	p value
chr3	131102053	NUDT16	3UTR;CDS	12	3.07	0.00	0.00	37.50	81.14	0
chr3	156259229	SSR3	3UTR	24	14.35	11.89	0.00	30.00	66.10	0
chr3	156259229	SSR3	3UTR	24	14.35	11.89	0.00	30.00	66.10	0
chr3	156259288	SSR3	3UTR	18	4.58	0.00	0.00	25.64	72.95	0
chr3	156259295	SSR3	3UTR	14	4.53	3.92	0.00	12.50	35.40	0.001
chr3	156259295	SSR3	3UTR	14	4.53	3.92	0.00	12.50	35.40	0.002
chr4	2940462	NOP14	3UTR	19	43.85	45.95	22.22	66.67	57.32	0
chr4	57326126	PAICS	3UTR	24	22.14	22.48	10.87	47.27	61.60	0
chr4	57326131	PAICS	3UTR	24	11.35	11.29	3.33	27.78	66.35	0
chr4	57326131	PAICS	3UTR	24	11.35	11.29	3.33	27.78	66.35	0
chr4	57326186	PAICS	3UTR	26	4.97	4.21	0.61	17.02	128.19	0
chr4	57326262	PAICS	3UTR	25	5.18	4.55	0.00	23.64	82.50	0
chr4	57326262	PAICS	3UTR	25	5.18	4.55	0.00	23.64	82.50	0
chr4	57326291	PAICS	3UTR	25	53.67	57.89	27.40	86.96	138.44	0
chr4	57326291	PAICS	3UTR	25	53.67	57.89	27.40	86.96	138.44	0
chr4	57326333	PAICS	3UTR	25	6.38	7.27	1.14	21.05	88.23	0
chr4	57326333	PAICS	3UTR	25	6.38	7.27	1.14	21.05	88.23	0
chr4	57326875	PAICS	3UTR	26	15.91	17.62	4.79	30.00	70.96	0
chr4	57326879	PAICS	3UTR	26	9.28	9.40	0.00	22.86	94.01	0
chr4	57326879	PAICS	3UTR	26	9.28	9.40	0.00	22.86	94.01	0
chr4	57326917	PAICS	3UTR	27	6.06	5.26	0.00	22.45	73.53	0
chr4	57327017	PAICS	3UTR	23	29.84	33.33	6.10	61.36	132.78	0
chr4	57327017	PAICS	3UTR	23	29.84	33.33	6.10	61.36	132.78	0
chr4	77979680	CCNI	CDS	24	6.96	4.97	0.00	23.53	137.09	0
chr4	183811670	DCTD	3UTR	26	20.75	0.00	0.00	100.00	1041.98	0
chr5	79923293	DHFR	3UTR	15	6.36	5.36	0.00	20.59	42.91	0
chr5	79923311	DHFR	3UTR	13	4.82	2.50	0.00	15.73	33.98	0.002
chr5	79923311	DHFR	3UTR	13	4.82	2.50	0.00	15.73	33.98	0.001
chr5	150639409	GM2A	CDS	23	44.00	46.15	0.00	100.00	1614.33	0
chr5	150639439	GM2A	CDS	23	43.44	34.43	0.00	100.00	1480.83	0
chr5	150647012	GM2A	3UTR; stop codon	26	24.89	33.67	0.00	100.00	1891.31	0
chr5	150648145	GM2A	3UTR	27	4.21	3.48	0.00	9.18	74.20	0
chr6	31238261	HLA-C	3UTR;CDS	19	0.12	0.06	0.00	3.23	51.39	0
chr6	34556786	C6orf106	3UTR	23	1.57	0.00	0.00	8.45	48.82	0.001

chrom	position	gene name	feature	n§	mean RDD level	median RDD level	minimum RDD level	maximum RDD level	Z‡	p value
chr6	41903007	CCND3	3UTR	14	46.15	45.24	0.00	100.00	397.77	0
chr6	42856330	RPL7L1	3UTR	19	16.10	15.15	0.89	35.48	96.00	0
chr6	153603584	intergenic	intergenic	27	8.19	0.06	0.00	99.53	30351.78	0
chr7	1976457	MAD1L1	CDS	18	20.13	0.00	0.00	100.00	591.13	0
chr8	11702542	CTSB	3UTR	21	34.72	44.44	0.00	100.00	693.10	0
chr8	17927327	ASAH1	CDS	24	50.94	45.49	0.00	100.00	1014.47	0
chr8	28206275	FBXO16	3UTR;CDS	22	27.98	0.00	0.00	100.00	882.59	0
chr8	30535953	GSR	3UTR	26	15.26	12.66	4.05	42.19	128.76	0
chr8	30535953	GSR	3UTR	26	15.26	12.66	4.05	42.19	128.76	0
chr8	30535980	GSR	3UTR	25	42.82	38.67	4.26	75.00	193.80	0
chr8	30535980	GSR	3UTR	25	42.82	38.67	4.26	75.00	193.80	0
chr8	30536016	GSR	3UTR	12	13.25	12.01	0.00	37.21	35.44	0
chr8	30536016	GSR	3UTR	12	13.25	12.01	0.00	37.21	35.44	0.002
chr8	30536078	GSR	3UTR	26	30.27	28.42	7.27	64.71	139.27	0
chr8	48889659	MCM4	3UTR	26	3.39	2.91	0.00	16.22	53.68	0.002
chr8	103841636	AZIN1	CDS	23	9.09	7.56	0.00	24.66	100.07	0
chr9	131071533	TRUB2	3UTR	23	10.37	10.42	0.00	23.33	49.05	0.003
chr9	132651866	FNBP1	3UTR	20	1.83	0.00	0.00	24.44	60.95	0
chr9	134373264	PRRC2B	3UTR	18	2.72	1.04	0.00	12.82	39.10	0.001
chrX	24095325	EIF2S3	3UTR	25	14.95	15.79	3.53	33.33	78.66	0

§ number of individuals with a minimum coverage of 30x at the RDD site

‡ test-statistic for evaluation of individual variation in RDD levels (see Materials and Methods)

Figure 5.2 Examples of sites showing significant variation in A-to-G RDD or editing levels among 27 unrelated individuals.



5.3.2 Evaluating the genetic component of individual variation in RDD levels

To analyze the genetic basis of variation in RDD levels, we obtained RNA-Seq data on 10 monozygotic (MZ) twin pairs and measured their RDD levels at sites in the DARNED database (see Materials and Methods). To be confident of the RDD levels, we restricted our analyses to RDD sites where at least 5 MZ twin pairs have a minimum coverage of 30x. In addition, we required a minimum of 10 samples with RDD levels greater than 0. In total, 10,993 sites met these criteria. For each of these sites, we assessed the genetic component to the variation in RDD levels (see Materials and Methods). Using a false positive rate of 0.01 (Figure 5.3), we identified 2,099 (19%) sites in which there is a significant genetic basis for variation in RDD levels (FDR ~ 0.05). These sites reside in approximately 500 different genes. A list of 50 top candidates is shown in Table 5.2. A few examples are shown in Figures 5.4 to 5.6.

Figure 5.3 False discovery rate versus false positive rate for evaluation of genetic basis of RDDs. Here we calculate the false discovery rate versus false positive rate after correcting for multiple testing error in the assessment of the genetic basis of RDDs in 10 pairs of MZ twins. In particular, the test statistic used here is an unequal variance ANOVA measurement (see Materials and Methods).

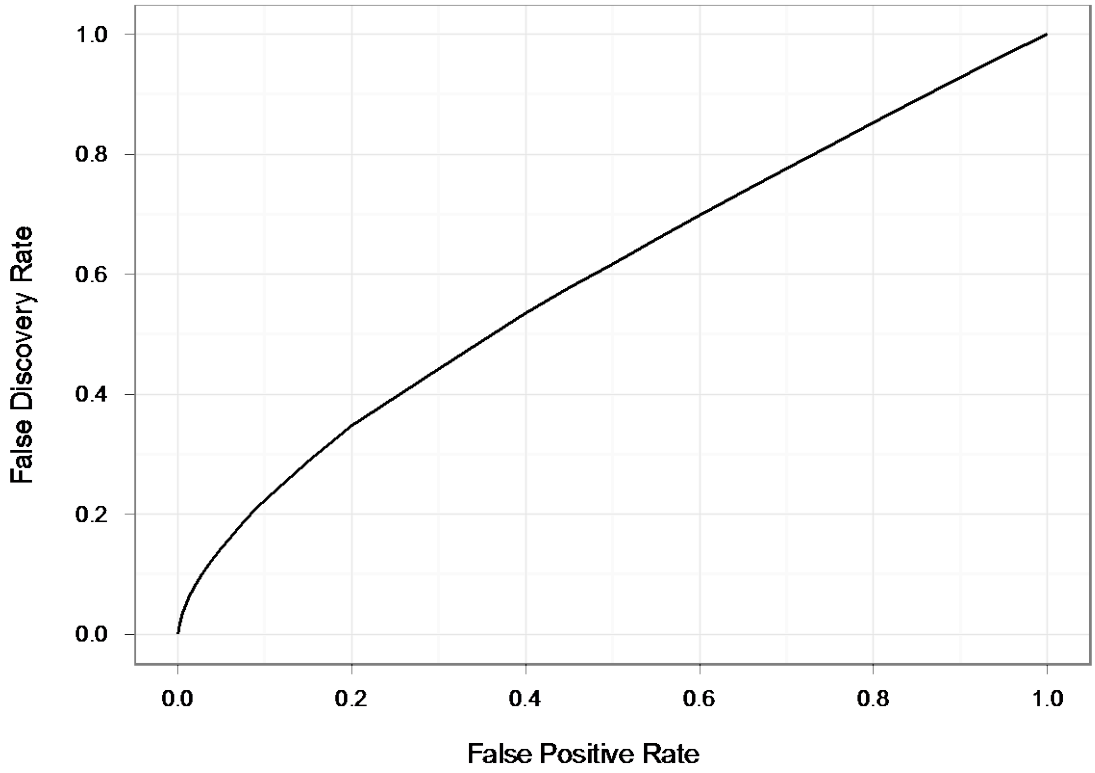


Table 5.2 List of top A-to-G sites with significant genetic component to individual variation in RNA editing levels among 10 monozygotic twin pairs.

chrom	end	gene	feature	# of pairs§	# of individuals‡	SS within	SS between	R^2	p value
chr1	1247494	CPSF3L	3UTR;CDS	20	19	2.49	778.49	1.00	0.00
chr1	1595074	SLC35E2B	3UTR	18	18	0.29	9.89	0.97	0.00
chr1	1595586	SLC35E2B	3UTR	20	20	1.90	18.84	0.91	0.00
chr1	1660887	SLC35E2	intron	12	12	0.63	28.13	0.98	0.00
chr1	6158562	KCNAB2	CDS	20	18	3.73	1065.60	1.00	0.00
chr1	6159032	KCNAB2	3UTR	20	20	0.41	281.44	1.00	0.00
chr1	6160876	KCNAB2	3UTR	20	20	0.61	362.67	1.00	0.00
chr1	6160958	KCNAB2	3UTR	20	20	0.10	476.89	1.00	0.00
chr1	36068370	PSMB2	3UTR	20	20	7.29	76.59	0.91	0.00
chr1	38327384	INPP5B	3UTR	20	16	3.20	157.81	0.98	0.00
chr1	45242978	RPS8	exon	14	14	0.12	6.78	0.98	0.00
chr1	46078630	NASP	exon	10	10	0.10	28.02	1.00	0.00
chr1	53291435	ZYG11B	3UTR	10	10	1.23	6.18	0.83	0.00
chr1	67874859	SERBP1	3UTR	20	20	5.79	50.94	0.90	0.00
chr1	144829580	NBPF9	3UTR	10	10	0.85	4.75	0.85	0.00
chr1	144829720	NBPF9	3UTR	20	20	3.12	36.62	0.92	0.00
chr1	150975108	FAM63A	5UTR;CDS	14	14	0.03	17.66	1.00	0.00
chr1	160966351	F11R	3UTR	20	20	5.14	49.13	0.91	0.00
chr1	160966352	F11R	3UTR	20	20	9.30	80.22	0.90	0.00
chr1	160967938	F11R	3UTR	20	20	9.44	79.53	0.89	0.00
chr1	184761274	FAM129A	3UTR	20	20	4.16	49.40	0.92	0.00
chr1	184761350	FAM129A	3UTR	16	16	5.00	32.25	0.87	0.00
chr1	184762524	FAM129A	3UTR	16	16	8.09	60.54	0.88	0.00
chr11	5719095	TRIM22	exon	20	14	0.24	264.99	1.00	0.00
chr11	65544382	AP5B1	3UTR	12	12	0.59	7.34	0.93	0.00

chrom	end	gene	feature	# of pairs§	# of individuals‡	SS within	SS between	R^2	p value
chr12	44181749	IRAK4	3UTR	10	10	0.70	5.71	0.89	0.00
chr12	72096125	TMEM19	3UTR	12	12	0.97	14.12	0.94	0.00
chr12	98942688	TMPO	3UTR	20	20	0.51	6.57	0.93	0.00
chr12	98942695	TMPO	3UTR	20	20	8.08	87.19	0.92	0.00
chr12	98943033	TMPO	3UTR	20	20	0.18	13.63	0.99	0.00
chr12	122216180	RHOF	3UTR	20	11	0.00	0.04	0.92	0.00
chr13	20247023	MPHOSPH8	3UTR	20	20	4.47	38.36	0.90	0.00
chr13	20247116	MPHOSPH8	3UTR	20	20	2.35	23.47	0.91	0.00
chr13	111547851	ANKRD10	intron	10	10	0.82	4.58	0.85	0.00
chr14	20916958	OSGEP	5UTR;CDS	20	18	1.02	566.37	1.00	0.00
chr15	41809530	RPAP1	3UTR	20	20	0.12	92.02	1.00	0.00
chr15	64447076	SNX22	3UTR	12	12	0.39	3.50	0.90	0.00
chr16	15795035	NDE1	intron	20	18	0.55	37.15	0.99	0.00
chr16	28967974	NFATC2IP	exon	12	12	1.08	20.93	0.95	0.00
chr16	29679989	SPN	3UTR	20	20	0.28	22.77	0.99	0.00
chr2	37327662	EIF2AK2	3UTR	20	20	1.93	44.96	0.96	0.00
chr2	37328034	EIF2AK2	3UTR	20	20	9.67	107.87	0.92	0.00
chr2	37328097	EIF2AK2	3UTR	20	20	24.44	251.67	0.91	0.00
chr2	99812336	MRPL30	3UTR	20	20	3.63	63.80	0.95	0.00
chr4	17803019	DCAF16	3UTR	20	20	1.24	11.01	0.90	0.00
chr4	57327058	PAICS	3UTR	20	20	3.18	85.12	0.96	0.00
chr4	73922926	COX18	3UTR	20	20	0.68	13.64	0.95	0.00
chr4	73922978	COX18	3UTR	20	20	1.76	25.46	0.94	0.00
chr4	73923097	COX18	3UTR	20	20	2.60	27.45	0.91	0.00
chr4	73923449	COX18	3UTR	20	20	2.32	22.10	0.90	0.00

§ number of MZ twin pairs where both members in pair have minimum coverage of 30x

‡ number of individuals with RDD levels greater than 0

Figure 5.4 A-to-G editing levels for 10 pairs of monozygotic twins in the 3' UTR of the gene F11R at chr1:160966352 (hg19)

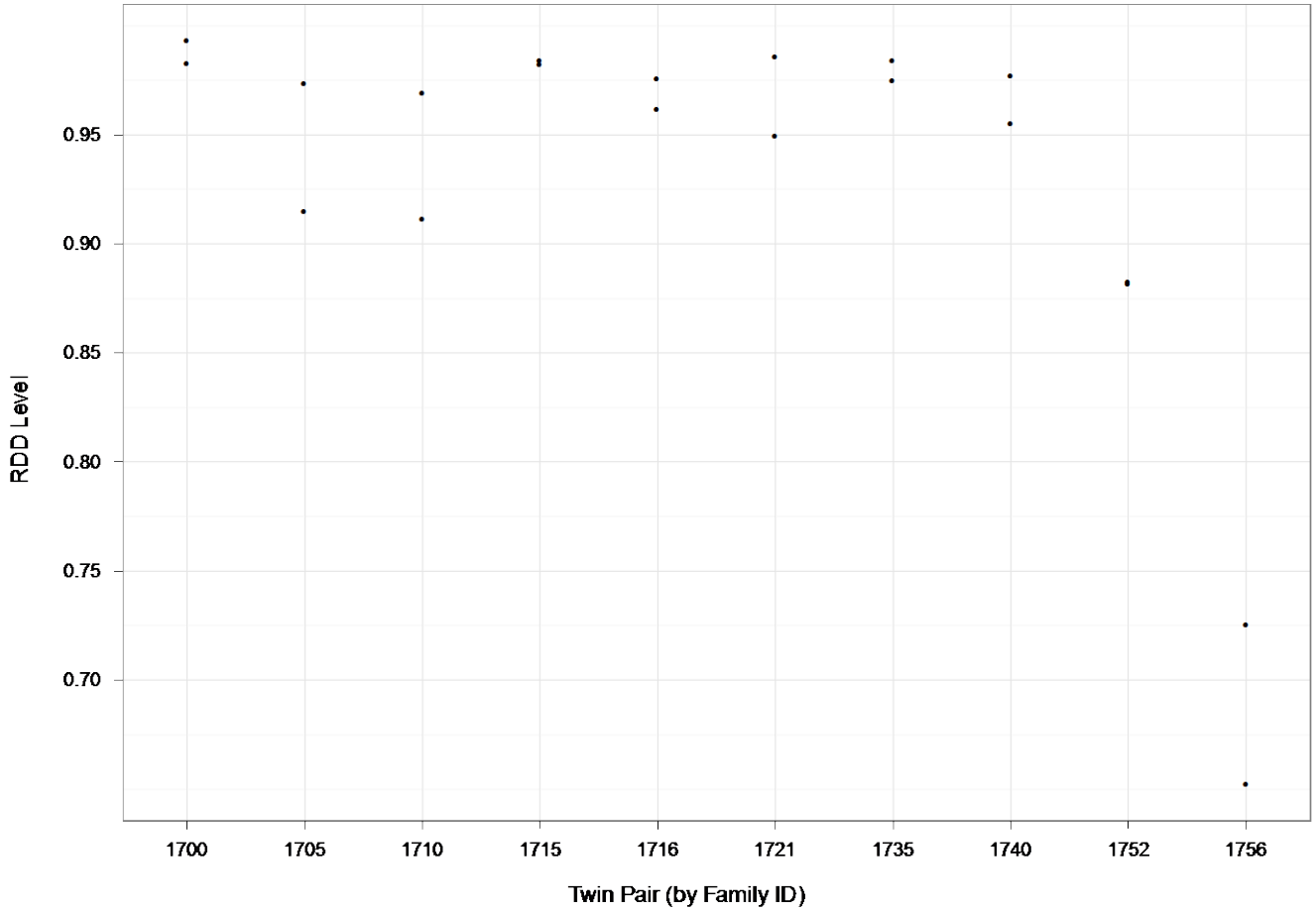


Figure 5.5 A-to-G editing levels for 10 pairs of monozygotic twins in the 3' UTR of the gene EIF2AK2 at chr2: 37327662 (hg19)

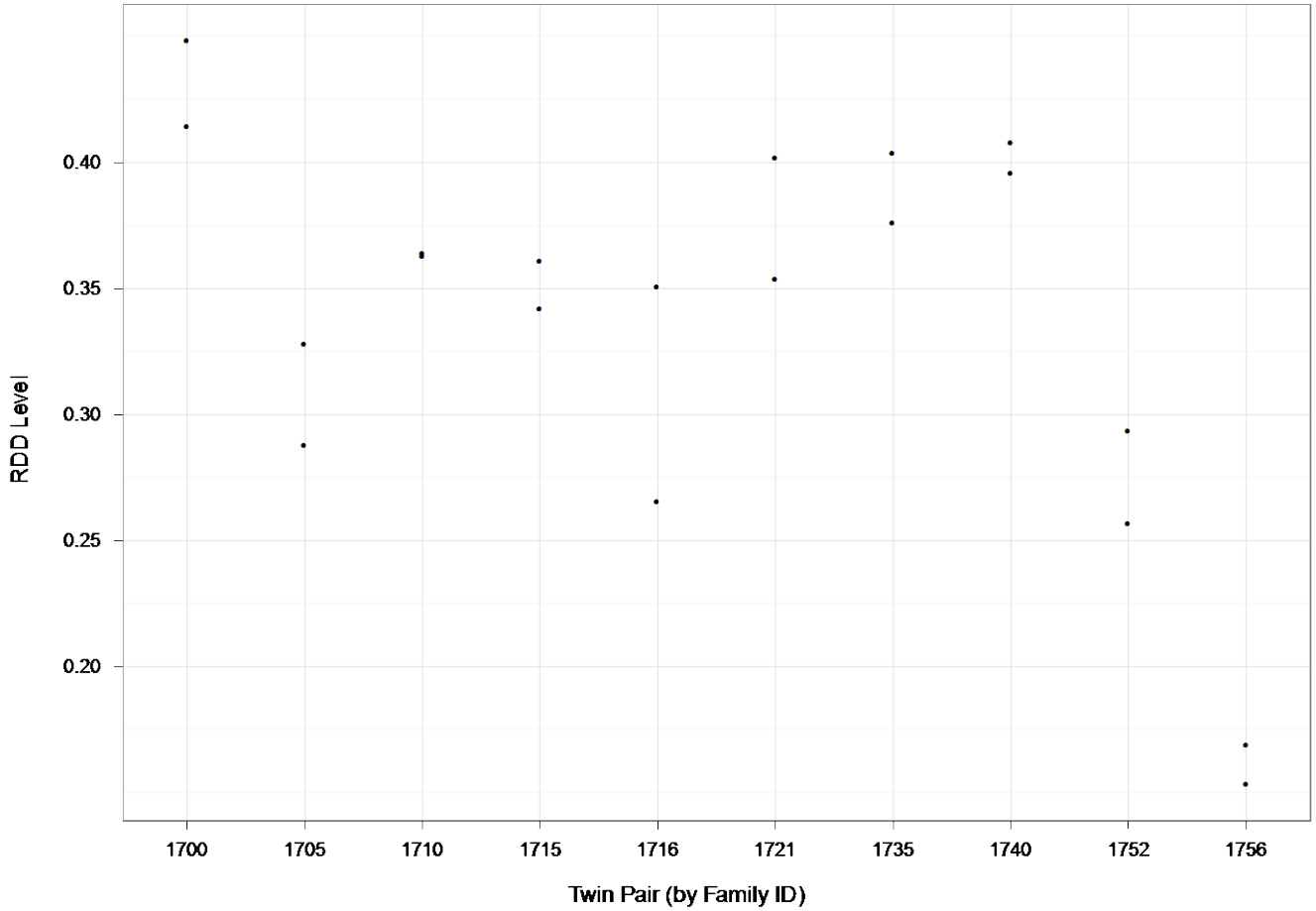
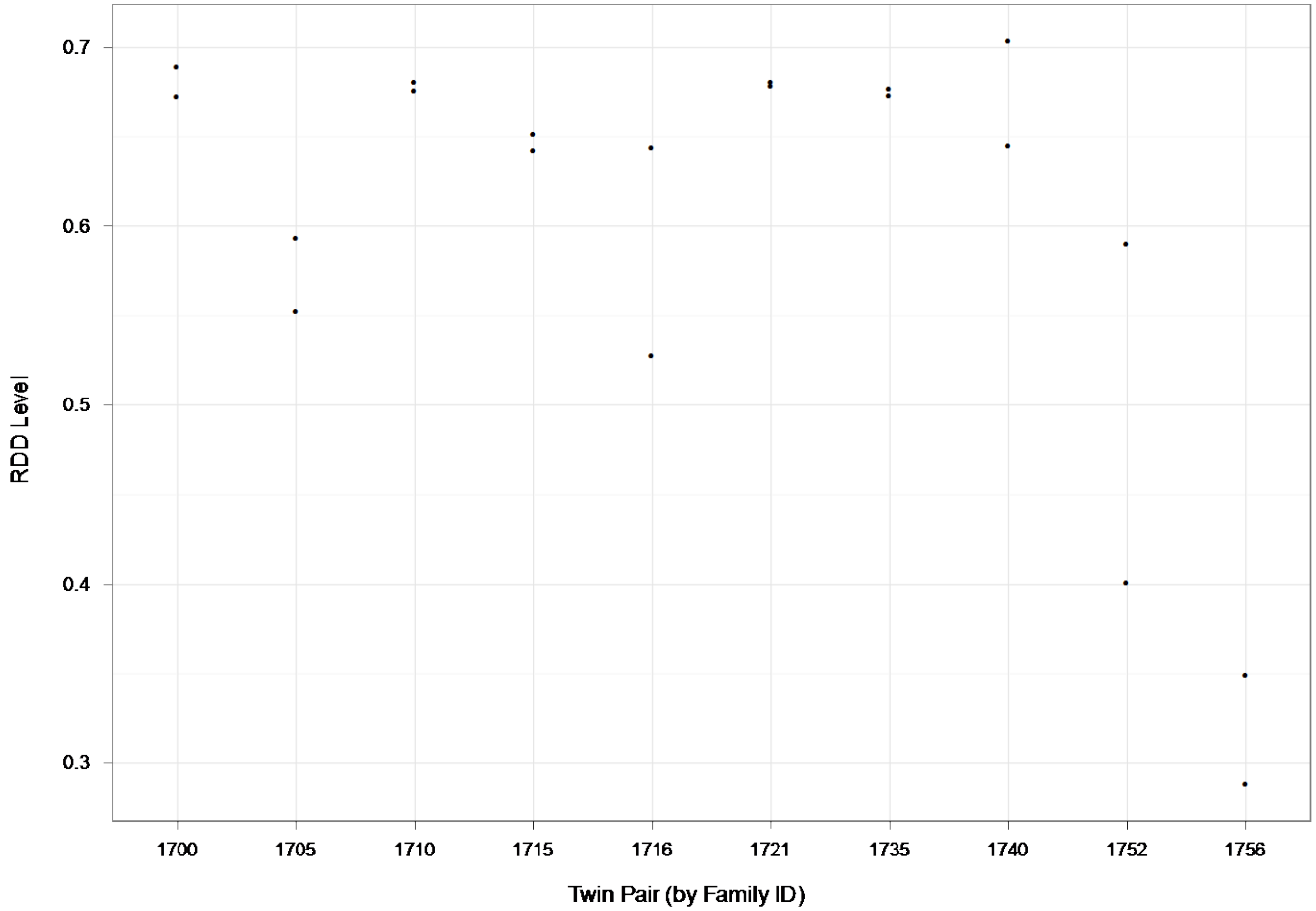


Figure 5.6 A-to-G editing levels for 10 pairs of monozygotic twins in the 3' UTR of the gene PAICS at chr4:57327058 (hg19)



5.4 Discussion and Future Directions

RNA-Seq provides digital and quantitative measurements of the transcriptome. The single-nucleotide resolution afforded by next-generation sequencing allows for detailed analyses of sequence differences between DNA and RNA and the precise measurement of the degree of these differences. In this chapter, we defined the RNA-DNA sequence difference (RDD) level, or the proportion of transcripts at an RDD site bearing the sequence difference allele, as a quantitative phenotype and investigated whether it is variable among individuals. Our work represents the study on individual variation of RDD levels at the genome-wide level, as previous reports focused only on a few genes. Using RNA-Seq data from 27 unrelated individuals, we detected over 100 sites in roughly 60 genes where the levels of A-to-G RNA editing among the samples vary more than expected under random sampling error. These samples were processed more than 5 years ago in 2008 and were thus sequenced to a relatively low depth of approximately 8x on average. We anticipate that with higher depths of sequencing and a larger set of individuals, we will observe a greater percentage of sites showing significant individual variation.

In addition to assessing the presence of variation in RDD levels among unrelated individuals, we also asked whether there is a genetic basis to such variation. We obtained RNA-Seq data on 10 pairs of monozygotic twins and measured their A-to-G editing levels across the genome. We found that nearly 20% of the 10,993 sites for which we had sufficient RNA-Seq coverage had a significant genetic component to the variation in RDD levels. Our results suggest that there is a hint of heritability for RDDs in humans. We note that while our dataset contains only 10 pairs of individuals, we have measured

the RDD level phenotype for thousands of sites. We expect that with deeper sequencing coverage and a larger set of twin pairs, we will observe a finer estimate of the heritability of RDDs.

At the time of this study, DNA-Seq data on the twin pairs was not available, and thus we did not explore the heritability of noncanonical RDDs. Future directions will involve expanding the analysis to RDDs of other types. In addition to exploring the heritability of RDDs for which we do not know the mechanism, we plan to understand the underlying polymorphisms that influence variation in A-to-G editing levels. We plan to identify single-nucleotide polymorphisms via DNA-Sequencing and perform a genome-wide association test for sites demonstrating a significant genetic component to the variation in RDD levels. We will also test for correlation between expression levels of ADAR and the editing levels of A-to-G sites.

In summary, our work lays the foundation for future genetic analyses of RDDs. Our study suggests that there exists a hint of heritability in A-to-G editing levels and future work will focus on (1) whether a genetic basis exists for RDDs of other types, (2) what are the exact genetic determinants that influence variation in A-to-G editing, (3) what are the functional consequences of differences in editing levels, and (4) linkage and association analyses of RDDs to uncover the mechanisms underlying noncanonical RDDs if they exist.

5.5 Materials and Methods

5.5.1 Samples

For the study of individual variation in unrelated individuals, we used 27 unrelated individuals from the Utah pedigrees of the Center d'Étude du Polymorphisme Humain collection (CEPH) as previously described (M. Li et al. 2011). For the study of the genetic basis of RDD levels, we used 10 pairs of monozygotic (MZ) twins from the CEPH database. See Table 5.3 for the list of individuals used in either dataset. The cultured B-cells were grown to a density of 5 x 10^5 cells/mL in RPMI 1640 supplemented with 15% fetal bovine serum, 100 units/mL penicillin and 100 μg/mL streptomycin, and 2 mmol/L L-glutamine.

5.5.2 RNA-Sequencing

RNA-Seq libraries were prepared as recommended by Illumina. Briefly, cells were harvested 24 hours after addition of fresh medium, and total RNA was extracted using the RNeasy Mini kit with DNase treatment (Qiagen). Poly-A mRNA was isolated and fragmented. First strand cDNA was prepared using reverse transcriptase and random hexamers. After second strand cDNA synthesis, ends were repaired and a single 'A' base was added followed by adapter sequences. Library fragments were selected for an average size of 200 to 260 base pairs, PCR amplified, and then sequenced using the Illumina Genome Analyzer 1, Illumina Genome Analyzer II, or HiSeq 2000 platform (Table 5.3).

Table 5.3 RNA-Sequencing statistics.

Sample	Family	System†	Number of Reads Sequenced	Read Length	Number of Reads Aligned	Number of Reads Aligned Uniquely	Mean Coverage	Median Coverage	Mean Coverage (Non Repeat Masker Regions)	Median Coverage (Non Repeat Masker Regions)
Unrelated Individuals (*n* = 27)										
GM06985	1341	1G GA	38,424,688	1x50bp	29,150,529 (76%)	22,275,984 (58%)	6	1	8	1
GM06994	1340	1G GA	37,634,150	1x50bp	30,408,512 (81%)	23,887,949 (63%)	9	1	12	2
GM07000	1340	1G GA	36,665,036	1x50bp	28,450,757 (78%)	21,124,653 (58%)	8	1	11	2
GM11829	1350	1G GA	37,293,182	1x50bp	27,229,474 (73%)	20,293,401 (54%)	12	1	15	2
GM11830	1350	1G GA	36,412,185	1x50bp	27,347,752 (75%)	20,445,547 (56%)	10	1	13	2
GM11831	1350	1G GA	38,923,882	1x50bp	27,163,969 (70%)	20,389,054 (52%)	8	1	11	2
GM11832	1350	1G GA	38,657,782	1x50bp	29,347,829 (76%)	22,612,842 (58%)	8	1	11	2
GM11881	1347	1G GA	49,771,575	1x50bp	31,903,474 (64%)	24,236,979 (49%)	7	1	9	1
GM11992	1362	1G GA	63,260,299	1x50bp	42,188,915 (67%)	30,404,083 (48%)	10	1	14	2
GM11993	1362	1G GA	38,441,615	1x50bp	29,827,543 (78%)	22,441,894 (58%)	8	1	10	1
GM11994	1362	1G GA	42,390,584	1x50bp	31,665,651 (75%)	23,910,152 (56%)	6	1	9	1
GM12003	1420	1G GA	36,977,676	1x50bp	25,506,005 (69%)	19,079,343 (52%)	12	2	15	2
GM12004	1420	1G GA	28,390,919	1x50bp	19,393,871 (68%)	14,275,512 (50%)	9	1	12	2
GM12005	1420	1G GA	37,359,318	1x50bp	25,389,881 (68%)	18,366,252 (49%)	5	1	8	1
GM12006	1420	1G GA	35,506,222	1x50bp	26,439,340 (74%)	19,972,831 (56%)	15	2	18	2
GM12043	1346	1G GA	36,999,366	1x50bp	26,110,944 (71%)	20,412,688 (55%)	7	1	10	1
GM12044	1346	1G GA	39,839,492	1x50bp	28,899,913 (73%)	21,490,915 (54%)	11	1	14	2
GM12144	1334	1G GA	35,020,229	1x50bp	24,663,047 (70%)	18,091,383 (52%)	6	1	8	1
GM12155	1408	1G GA	43,551,373	1x50bp	32,059,769 (74%)	24,429,087 (56%)	12	1	16	2
GM12716	1358	1G GA	46,656,852	1x50bp	34,787,074 (75%)	26,301,181 (56%)	6	1	9	1
GM12717	1358	1G GA	40,635,779	1x50bp	31,736,582 (78%)	24,788,870 (61%)	12	1	15	2
GM12750	1444	1G GA	45,066,631	1x50bp	32,989,463 (73%)	25,354,509 (56%)	9	1	12	1
GM12762	1447	1G GA	47,651,198	1x50bp	33,962,134 (71%)	25,759,636 (54%)	8	1	11	1
GM12813	1454	1G GA	45,288,662	1x50bp	35,704,677 (79%)	27,765,787 (61%)	10	1	14	2
GM12814	1454	1G GA	39,024,182	1x50bp	29,749,717 (76%)	22,620,609 (58%)	10	1	13	2
GM12872	1459	1G GA	45,367,154	1x50bp	29,379,313 (65%)	22,143,485 (49%)	5	1	7	2
GM12874	1459	1G GA	39,588,081	1x50bp	26,863,489 (68%)	19,637,631 (50%)	6	1	8	1

Sample	Family	System†	Number of Reads Sequenced	Read Length	Number of Reads Aligned	Number of Reads Aligned Uniquely	Mean Coverage	Median Coverage	Mean Coverage (Non Repeat Masker Regions)	Median Coverage (Non Repeat Masker Regions)
MZ Twins (*n* = 20)										
GM14381	1700	HiSeq 2000	87,747,958	1x100bp	85,611,174 (98%)	72,682,214 (83%)	25	2	36	2
GM14382	1700	HiSeq 2000	96,532,207	1x100bp	94,112,639 (97%)	79,605,763 (82%)	26	2	38	2
GM14408	1705	HiSeq 2000	190,370,877	1x100bp	186,669,448 (98%)	160,069,233 (84%)	45	2	68	3
GM14409	1705	HiSeq 2000	192,128,604	1x100bp	187,853,801 (98%)	159,841,274 (83%)	41	2	62	3
GM14432	1710	HiSeq 2000	99,374,247	1x100bp	95,736,606 (96%)	82,128,151 (83%)	25	2	36	3
GM14433	1710	HiSeq 2000	109,389,141	1x100bp	106,549,818 (97%)	91,049,099 (83%)	26	2	38	3
GM14447	1715	HiSeq 2000	88,964,274	1x100bp	87,397,953 (98%)	74,813,313 (84%)	23	2	33	3
GM14448	1715	HiSeq 2000	94,936,103	1x100bp	92,579,625 (98%)	79,541,620 (84%)	18	2	28	2
GM14452	1716	HiSeq 2000	107,194,377	1x100bp	105,233,707 (98%)	87,869,309 (82%)	25	2	36	3
GM14453	1716	HiSeq 2000	121,015,870	1x100bp	118,648,218 (98%)	102,374,991 (85%)	33	2	48	2
GM14467	1721	HiSeq 2000	112,512,422	1x100bp	109,872,178 (98%)	93,517,681 (83%)	26	2	38	2
GM14468	1721	HiSeq 2000	130,482,667	1x100bp	127,008,691 (97%)	106,994,341 (82%)	24	2	36	2
GM14506	1735	HiSeq 2000	100,444,954	1x100bp	98,394,404 (98%)	81,894,039 (82%)	27	2	40	2
GM14507	1735	HiSeq 2000	104,762,561	1x100bp	102,466,902 (98%)	86,322,550 (82%)	28	2	41	3
GM14520	1740	HiSeq 2000	105,916,302	1x100bp	103,691,234 (98%)	85,923,974 (81%)	29	2	43	2
GM14521	1740	HiSeq 2000	120,920,893	1x100bp	118,644,112 (98%)	99,258,597 (82%)	32	2	46	2
GM14568	1752	HiSeq 2000	109,660,421	1x100bp	107,666,287 (98%)	92,728,585 (85%)	28	2	41	2
GM14569	1752	HiSeq 2000	122,436,596	1x100bp	120,514,178 (98%)	103,772,291 (85%)	18	2	28	2
GM14581	1756	HiSeq 2000	189,630,442	1x100bp	186,763,322 (98%)	156,446,506 (83%)	42	2	63	3
GM14582	1756	HiSeq 2000	170,152,070	1x100bp	167,651,625 (99%)	139,468,387 (82%)	41	2	62	3

† System: 1G GA = Illumina Genome Analyzer; 2G GA = Illumina Genome Analyzer 2

5.5.3 Alignment of RNA-Seq datasets

RNA-Seq datasets were aligned using GSNAP version 2012-07-20 (T. D. Wu & Nacu 2010) to the human genome (build hg19). GSNAP was run with default options. A maximum number of 10 alignments were permitted for each read. Alignments to novel exon-exon junctions (per GSNAP option -N 1) and known junctions as defined by RefSeq (downloaded November 2, 2012) and Gencode version 13 (Harrow et al. 2006) were accepted. Alignments with no more than the default maximum of '(read length + 2)/12 – 2' mismatches were retained. Non-primary alignments and alignments placing read pairs in the incorrect orientation were removed from downstream analyses.

5.5.4 Assessment of individual variation in RNA-DNA sequence difference levels

To quantify the significance of variation in RDD levels among unrelated individuals, we developed a formal statistical test. In particular, we assumed that RDD levels follow a binomial distribution as follows:

$$y_i \sim binomial(n_i, p_i)$$

where y_i is the number of reads bearing the RDD base,
n_i is the coverage at the site, and
p_i is RDD level for individual i

We defined the null and alternative hypotheses as follows:

$$H_0: \quad p_1 = p_1 = p_1 = \cdots = p_m = p$$

$$H_A: \quad \text{at least one } p_i \text{ differs}$$

Under the null hypothesis, the estimate for p is the average across the individuals, or:

$$\hat{p} = \frac{y_1 + y_2 + \cdots + y_m}{n_1 + n_2 + \cdots + n_m}$$

We calculate the log-likelihood as the following:

$$L_0 = \log \prod_{i=1}^{m} [\hat{p}^{y_i}(1-\hat{p})^{n-y_i}] \binom{n_i}{y_i}$$

$$L_A = \log \prod_{i=1}^{m} [\hat{p}_i{}^{y_i}(1-\hat{p}_i)^{n-y_i}] \binom{n_i}{y_i}$$

We define our test statistic Z as $2(L_A - L_0)$, or:

$$Z \triangleq \sum_{i=1}^{n} \left[y_i \log\left(\frac{\hat{p}_i}{\hat{p}}\right) + (n_i - y_i) \log\left(\frac{1-\hat{p}_i}{1-\hat{p}}\right) \right]$$

We observe that our test statistic Z is large if the variation in RDD levels is more than what is expected under random sampling error.

To correct for multiple testing error, we used bootstrapping to estimate the sampling distribution. In particular, for each iteration j, we generated new numbers of reads bearing the RDD base y^* for each individual m such that:

$$y_m^* \sim binomial(n_m, \hat{p})$$

For each iteration j of new RDD levels, we calculated a test statistic Z^{*j}. We rejected H_0, or claimed that there is significant individual variation in RDD levels at this site, if our value of Z derived from the observed RDD levels is greater than the $(1 - \alpha)^{th}$ percentile of Z^*, where α is the false positive rate. We then calculated the false discovery rate as the following:

$$FDR = \frac{\text{number of expected false positives}}{\text{number of RDD sites rejected}} \leq \frac{N * \alpha}{\text{number of RDD sites rejected}}$$

where N is the total number of RDD positions tested.

5.5.5 Assessment of genetic component to individual variation in RDD levels

To measure the component of individual variation in RDD levels that is explained by genetic factors, we developed a formal statistical test based on ANOVA and the intraclass correlation coefficient. First we define the RDD level p_{ij} as the following:

$$p_{ij} = \frac{y_{ij}}{n_{ij}}$$

where y_{ij} is the number of reads bearing the RDD base,
n_{ij} is the coverage at the site, and
p_{ij} is RDD level for twin pair *i* and individual *j*

Next, we define the average RDD level for each twin pair i as

$$\bar{p_i} = \frac{p_{i1} + p_{i2}}{2}$$

and the overall average RDD level across all individuals in the *n* twin pair dataset as

$$\bar{\bar{p}} = \sum_{i=1}^{n} \frac{\bar{p_i}}{n}$$

We assume that the RDD level p_{ij} for each individual follows the model

$$p_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where μ is the population RDD level,
α_i is the twin pair specific deviation
and ε_i is the error term.

We assume that α_i is normally distributed with a mean of 0 and variance of σ^2_A.

The total sum of squares SST is

$$SST = \sum_{i=1}^{n} (p_{i1} - \bar{\bar{p}})^2 + (p_{i2} - \bar{\bar{p}})^2$$

The within groups (or twin pairs) sum of squares is

$$SSWG = \sum_{i=1}^{n} (p_{i1} - \bar{p}_i)^2 + (p_{i2} - \bar{p}_i)^2$$

And lastly, the between groups sum of squares is

$$SSBG = SST - SSWG = \sum_{i=1}^{n} 2(\bar{p}_i - \bar{\bar{p}})^2$$

The coefficient of determination R^2 is

$$R^2 = \frac{SSBG}{SST}$$

Differences in sequencing depths across individuals will lead to variation in the sampling confidence of the true RDD level as sites with higher coverage have lower sampling error in RDD levels. To account such variation, we assumed that σ_{ij}^2 is the variance of the RDD level p_{ij} and that σ_{ij}^2 is known up to a constant or

$$\sigma_{ij}^2 \propto \frac{1}{n_{ij}}$$

In light of these differences in coverage, we formulate the sum of squares measurements as the following:

$$SST' = \sum_{i=1}^{n} \frac{(p_{i1} - \bar{\bar{p}})^2}{\sigma_{i1}^2} + \frac{(p_{i2} - \bar{\bar{p}})^2}{\sigma_{i2}^2}$$

Furthermore, we redefine the average RDD level for twin pair *i* and the within groups and between groups sum of squares as the following:

$$\bar{p}_i = \frac{\sigma_{i2}^2 p_{i1} + \sigma_{i1}^2 p_{i2}}{\sigma_{i1}^2 + \sigma_{i2}^2}$$

$$SSWG' = \sum_{i=1}^{n} \frac{(Y_{i1} - \bar{Y}_i)^2}{\sigma_{i1}^2} + \frac{(Y_{i2} - \bar{Y}_i)^2}{\sigma_{i2}^2}$$

$$SSBG' = \sum_{i=1}^{n} \frac{\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2}{\sigma_{i1}^2} + \frac{\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2}{\sigma_{i2}^2}$$

The coefficient of determination in an unequal (but known up to a constant) variance ANOVA is then

$$R^{2'} = \frac{SSBG'}{SST'}$$

To correct for multiple testing error, we generated random permutations of the dataset. For each permutation, twin pairs are reassigned such that each pair of individuals is are unrelated. For each iteration *j* of these new random assignments, we calculated a test statistic $R_j^{2'*}$. We claimed that there is significant genetic component to the variation in RDD levels at this site, if our value of $R^{2'}$ derived from the true assignment of twin pairs is greater than the $(1 - \alpha)^{\text{th}}$ percentile of $R^{2'*}$, where α is the false positive rate. We then calculated the false discovery rate as the following:

$$FDR = \frac{\text{number of expected false positives}}{\text{number of RDD sites rejected}} \leq \frac{N * \alpha}{\text{number of RDD sites rejected}}$$

where *N* is the total number of RDD positions tested.

Chapter 6. Conclusion

6.1 Summary of Work and Future Directions

With the completion of the human genome sequence, the next challenge for the genomics community lies in the proper annotation of functional elements and transcribed regions in the genome. Continual advances in sequencing technology have facilitated such research, enabling unprecedented interrogation of DNA and RNA sequence at the single-nucleotide level. These advances have spurred progress toward complete characterization of all transcribed elements in a population of cells, increasing the catalogue of known genes. In particular, the single-nucleotide resolution afforded by RNA-Seq has allowed for detailed analysis of alternative transcription start sites, alternative polyadenylation, alternative splicing, and RNA editing, and led as well to the discovery of sequence differences between RNA and DNA that cannot be explained by known mechanisms.

In chapter 2, we explored the use of next-generation sequencing technology to study the expression landscape in humans. We sequenced complementary DNA fragments derived from the RNA of human B-cell cell lines and obtained over 800 million reads comprising roughly 40 Gigabases of sequence. The abundance of sequence information allowed us to analyze the expression profile of the human transcriptome in depth and evaluate the experimental parameters necessary for sequencing-based studies. We measured the expression of 20,766 genes and 67,453 of their alternatively-spliced transcripts. We found that the vast majority (90%) of genes with multiple exons undergo alternative splicing. Our analyses of the appropriate sequencing depths for most transcriptome studies revealed that while a depth of 100 million reads is sufficient to

detect the presence of genes, at least 500 million reads are required to accurately measure their expression levels. Since the time this study began in 2008, advances in sequencing technology have lowered the cost of sequencing and greatly increased the throughput of a sequencing run. As the issue of attaining sequencing information becomes moot, greater emphasis will be placed on how to extract meaningful conclusions from the abundance of expression information on the transcriptome. Efforts to detect differential gene expression between various conditions (Robinson et al. 2010) and the ability to identify lowly expressed but functionally important genes (Halvardson et al. 2013; Mercer et al. 2012) remain top priorities.

Beyond the detection and quantification of genes, RNA-Seq technology allows for detailed comparisons of sequence. In particular, next-generation sequencing technology has paved the way for various studies focused on systematic sequence differences between DNA and RNA. In addition to A-to-G and C-to-T sequence differences in RNA as catalyzed by RNA editing processes, various researchers identified all 12 types of RNA-DNA sequence differences (RDDs), the majority of which cannot be explained by known mechanisms. In chapter 3, we describe an initial study performed in the Cheung laboratory in which comparisons of RNA and DNA sequences of 27 unrelated individuals uncovered over 10,000 exonic sites containing sequence differences between DNA and RNA. These RDDs were presumably non-random as many sites were found in different individuals and cell types.

The discovery of these noncanonical RDDs was shortly met thereafter with claims that technical artifacts explain the majority of these sequence differences. In chapter 4, we developed a detection theory approach for the optimal thresholds and parameters for

minimizing the false discovery rate of RDD identification. In particular, we generated synthetic datasets containing RDDs at known locations and evaluated the sensitivity and specificity of RDD detection using various aligners and within different regions of the genome. We also identified the major sources contributing to false positive RDDs and found that nonrandom sequencing errors present in experimental RNA-Seq datasets generated by Illumina sequencing technology play a major part in the false discovery rate of RDD detection. Moreover, we used our pipeline for the analysis of RDDs in a human cell line for which DNA and RNA data is readily available. We identified over 9,000 RDD events, the majority of which are A-to-G differences. The most common noncanonical RDD types we found were A-to-C and T-to-G differences, which have been shown to be the most common sequencing error. Furthermore, we found that the alignment quality for noncanonical RDD sites were in general of lower quality than that of A-to-G sites. Overall, we concluded that the evidence for widespread noncanonical RDDs in humans is weak.

RDD events can be viewed as a quantitative attribute that lends itself to genetic analyses. In chapter 5, we explored the genetic basis for RDDs by defining the RDD level, or proportion of reads at an RDD site containing the sequence difference allele, as a quantitative phenotype. In particular, we developed statistical methods for analyzing the variation in RDD levels among unrelated individuals, accounting for differences in sequencing depths and multiple testing error. We applied our algorithms to a real dataset comprising RNA-Seq data for 27 unrelated individuals. Overall, we found that nearly 10% of the sites for which we had sufficient coverage showed significant variation in RDD levels among samples. Previous studies on RNA editing levels in glutamate

receptors in contrast found consistent levels of editing across individuals. Future work will involve understanding the functional consequence of such variation.

In addition to asking whether RDD levels vary among unrelated individuals, we assessed whether such variation is explained by genetic factors. We formulated statistical tests for measuring the genetic component to variation in RDD levels using RNA-Seq data. In particular, we used an unequal variance ANOVA approach and accounted for variation in RNA-Seq coverage. We tested our algorithm on a RNA-Seq dataset derived from 10 monozygotic twin pairs. We found that nearly 20% of the sites with sufficient coverage had a significant genetic basis to the variation in RDD levels. Future directions include uncovering the specific cis- or trans-factors that influence such variation.

While many reports have challenged the initial discovery of RDDs by Cheung and colleagues, it is still unclear whether noncanonical sequence differences exist in humans. Most studies have focused on the high false discovery rates associated with studying RDDs using next-generation sequencing technology and not the identification of singular sites. If noncanonical RDDs do exist, the signal to noise ratio may be too low to allow detection in a genome-wide manner. Recent studies, however, have found previously unknown modifications in the transcriptome (Schaefer et al. 2009; Song et al. 2012). It will be interesting to see whether these differences overlap RDD sites.

In conclusion, next-generation sequencing technology is a powerful tool for examining the transcriptome in great detail. It provides a detailed and quantitative account of expression profiles, facilitating global studies of gene expression as well as specific analyses of sequence differences. RNA-Seq has paved the way for discovery into unchartered waters, and great restraint and discipline is required for proper analysis of

such novel and exciting data. We have entered an age where data collection has far outpaced data analysis, and now more than ever, do we need to develop methods for lucid and accurate interpretation of the genome.

Bibliography

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

Albers, C. A. *et al.* Dindel: accurate indel calls from short-read data. *Genome research* **21**, 961-973, doi:10.1101/gr.112326.110 (2011).

Alon, S. *et al.* Systematic identification of edited microRNAs in the human brain. *Genome research* **22**, 1533-1540, doi:10.1101/gr.131573.111 (2012).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).

Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29, doi:10.1038/75556 (2000).

Athanasiadis, A., Rich, A. & Maas, S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS biology* **2**, e391, doi:10.1371/journal.pbio.0020391 (2004).

Au, K. F., Jiang, H., Lin, L., Xing, Y. & Wong, W. H. Detection of splice junctions from paired-end RNA-Seq data by SpliceMap. *Nucleic acids research* **38**, 4570-4578, doi:10.1093/nar/gkq211 (2010).

Backus, J. W. & Smith, H. C. Specific 3' sequences flanking a minimal apolipoprotein B (apoB) mRNA editing 'cassette' are critical for efficient editing in vitro. *Biochimica et biophysica acta* **1217**, 65-73 (1994).

Bahn, J. H. *et al.* Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome research* **22**, 142-150, doi:10.1101/gr.124107.111 (2012).

Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003-1007, doi:10.1126/science.1072047 (2002).

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome research* **11**, 1005-1017, doi:10.1101/gr.187101 (2001).

Balzer, S., Malde, K., Lanzen, A., Sharma, A. & Jonassen, I. Characteristics of 454 pyrosequencing data--enabling realistic simulation with flowsim. *Bioinformatics* **26**, i420-425, doi:10.1093/bioinformatics/btq365 (2010).

Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837, doi:10.1016/j.cell.2007.05.009 (2007).

Bass, B. L. RNA editing by adenosine deaminases that act on RNA. *Annual review of biochemistry* **71**, 817-846, doi:10.1146/annurev.biochem.71.110601.135501 (2002).

Bass, B. L. & Weintraub, H. An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell* **55**, 1089-1098 (1988).

Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nature reviews. Genetics* **3**, 370-379, doi:10.1038/nrg798 (2002).

Beer, L. A., Tang, H. Y., Sriswasdi, S., Barnhart, K. T. & Speicher, D. W. Systematic discovery of ectopic pregnancy serum biomarkers using 3-D protein profiling coupled with label-free quantitation. *Journal of proteome research* **10**, 1126-1138, doi:10.1021/pr1008866 (2011).

Benne, R. *et al.* Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46**, 819-826 (1986).

Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59, doi:10.1038/nature07517 (2008).

Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242-2246, doi:10.1126/science.1103388 (2004).

Bhogal, B. *et al.* Modulation of dADAR-dependent RNA editing by the Drosophila fragile X mental retardation protein. *Nature neuroscience* **14**, 1517-1524, doi:10.1038/nn.2950 (2011).

Bianchetti, L., Kieffer, D., Federkeil, R. & Poch, O. Increased frequency of single base substitutions in a population of transcripts expressed in cancer cells. *BMC cancer* **12**, 509, doi:10.1186/1471-2407-12-509 (2012).

Bittner, M. *et al.* Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536-540, doi:10.1038/35020115 (2000).

Blank, A., Gallant, J. A., Burgess, R. R. & Loeb, L. A. An RNA polymerase mutant with reduced accuracy of chain elongation. *Biochemistry* **25**, 5920-5928 (1986).

Bock, R., Hermann, M. & Kossel, H. In vivo dissection of cis-acting determinants for plastid RNA editing. *The EMBO journal* **15**, 5052-5059 (1996).

Bock, R., Kossel, H. & Maliga, P. Introduction of a heterologous editing site into the tobacco plastid genome: the lack of RNA editing leads to a mutant phenotype. *The EMBO journal* **13**, 4623-4628 (1994).

Boguski, M. S., Tolstoshev, C. M. & Bassett, D. E., Jr. Gene discovery in dbEST. *Science* **265**, 1993-1994 (1994).

Borchert, G. M. *et al.* Adenosine deamination in human transcripts generates novel microRNA binding sites. *Human molecular genetics* **18**, 4801-4807, doi:10.1093/hmg/ddp443 (2009).

Borukhov, S., Sagitov, V. & Goldfarb, A. Transcript cleavage factors from E. coli. *Cell* **72**, 459-466 (1993).

Bourara, K., Litvak, S. & Araya, A. Generation of G-to-A and C-to-U changes in HIV-1 transcripts by RNA editing. *Science* **289**, 1564-1566 (2000).

Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-322, doi:10.1016/j.cell.2007.12.014 (2008).

Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S. R. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 3960-3964, doi:10.1073/pnas.0230489100 (2003).

Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752-755, doi:10.1126/science.1069516 (2002).

Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology* **18**, 630-634, doi:10.1038/76469 (2000).

Brockman, W. *et al.* Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome research* **18**, 763-770, doi:10.1101/gr.070227.107 (2008).

Brueckner, F., Hennecke, U., Carell, T. & Cramer, P. CPD damage recognition by transcribing RNA polymerase II. *Science* **315**, 859-862, doi:10.1126/science.1135400 (2007).

Brulliard, M. *et al.* Nonrandom variations in human cancer ESTs indicate that mRNA heterogeneity increases during carcinogenesis. *Proceedings of the National*

Academy of Sciences of the United States of America **104**, 7522-7527, doi:10.1073/pnas.0611076104 (2007).

Bruno, V. M. *et al.* Comprehensive annotation of the transcriptome of the human fungal pathogen Candida albicans using RNA-seq. *Genome research* **20**, 1451-1458, doi:10.1101/gr.109553.110 (2010).

Brusa, R. *et al.* Early-onset epilepsy and postnatal lethality associated with an editing-deficient GluR-B allele in mice. *Science* **270**, 1677-1680 (1995).

Bryant, D. W., Jr., Wong, W. K. & Mockler, T. C. QSRA: a quality-value guided de novo short read assembler. *BMC bioinformatics* **10**, 69, doi:10.1186/1471-2105-10-69 (2009).

Burger, G., Yan, Y., Javadi, P. & Lang, B. F. Group I-intron trans-splicing and mRNA editing in the mitochondria of placozoan animals. *Trends in genetics : TIG* **25**, 381-386, doi:10.1016/j.tig.2009.07.003 (2009).

Burns, C. M. *et al.* Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature* **387**, 303-308, doi:10.1038/387303a0 (1997).

Burrows, M. & Wheeler, D. A block-sorting lossless data compression algorithm. *SRC Research Reports* **124** (1994).

Butler, J. *et al.* ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research* **18**, 810-820, doi:10.1101/gr.7337908 (2008).

Campagna, D. *et al.* PASS: a program to align short sequences. *Bioinformatics* **25**, 967-968, doi:10.1093/bioinformatics/btp087 (2009).

Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics* **40**, 722-729, doi:10.1038/ng.128 (2008).

Cann, H. M. CEPH maps. *Current opinion in genetics & development* **2**, 393-399 (1992).

Cattaneo, R., Kaelin, K., Baczko, K. & Billeter, M. A. Measles virus editing provides an additional cysteine-rich protein. *Cell* **56**, 759-764 (1989).

Chaisson, M. J., Brinza, D. & Pevzner, P. A. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome research* **19**, 336-346, doi:10.1101/gr.079053.108 (2009).

Chaisson, M. J. & Pevzner, P. A. Short read fragment assembly of bacterial genomes. *Genome research* **18**, 324-330, doi:10.1101/gr.7088808 (2008).

Chan, L. Apolipoprotein B, the major protein component of triglyceride-rich and low density lipoproteins. *The Journal of biological chemistry* **267**, 25621-25624 (1992).

Chaudhuri, S., Carrer, H. & Maliga, P. Site-specific factor involved in the editing of the psbL mRNA in tobacco plastids. *The EMBO journal* **14**, 2951-2957 (1995).

Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods* **6**, 677-681, doi:10.1038/nmeth.1363 (2009).

Chen, S. H. *et al.* Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science* **238**, 363-366 (1987).

Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-1117, doi:10.1016/j.cell.2008.04.043 (2008).

Chepelev, I. Detection of RNA editing events in human cells using high-throughput sequencing. *Methods Mol Biol* **815**, 91-102, doi:10.1007/978-1-61779-424-7_8 (2012).

Chester, A., Scott, J., Anant, S. & Navaratnam, N. RNA editing: cytidine to uridine conversion in apolipoprotein B mRNA. *Biochimica et biophysica acta* **1494**, 1-13 (2000).

Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods* **6**, 99-103, doi:10.1038/nmeth.1276 (2009).

Cho, R. J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular cell* **2**, 65-73 (1998).

Cho, R. J. *et al.* Transcriptional regulation and function during the human cell cycle. *Nature genetics* **27**, 48-54, doi:10.1038/83751 (2001).

Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature methods* **5**, 613-619, doi:10.1038/nmeth.1223 (2008).

Cocquet, J., Chong, A., Zhang, G. & Veitia, R. A. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88**, 127-131, doi:10.1016/j.ygeno.2005.12.013 (2006).

Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215-219, doi:10.1038/nature06745 (2008).

Conticello, S. G. The AID/APOBEC family of nucleic acid mutators. *Genome biology* **9**, 229, doi:10.1186/gb-2008-9-6-229 (2008).

Conticello, S. G. Creative deaminases, self-inflicted damage, and genome evolution. *Annals of the New York Academy of Sciences* **1267**, 79-85, doi:10.1111/j.1749-6632.2012.06614.x (2012).

Cooper, T. A., Wan, L. & Dreyfuss, G. RNA and disease. *Cell* **136**, 777-793, doi:10.1016/j.cell.2009.02.011 (2009).

Covello, P. S. & Gray, M. W. RNA editing in plant mitochondria. *Nature* **341**, 662-666, doi:10.1038/341662a0 (1989).

Crawford, J. E. *et al.* De novo transcriptome sequencing in Anopheles funestus using Illumina RNA-Seq technology. *PloS one* **5**, e14202, doi:10.1371/journal.pone.0014202 (2010).

Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).

Damsma, G. E., Alt, A., Brueckner, F., Carell, T. & Cramer, P. Mechanism of transcriptional stalling at cisplatin-damaged DNA. *Nature structural & molecular biology* **14**, 1127-1133, doi:10.1038/nsmb1314 (2007).

Dausset, J. *et al.* Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6**, 575-577 (1990).

Davuluri, R. V., Suzuki, Y., Sugano, S., Plass, C. & Huang, T. H. The functional consequences of alternative promoter use in mammalian genomes. *Trends in genetics : TIG* **24**, 167-177, doi:10.1016/j.tig.2008.01.008 (2008).

De Bona, F., Ossowski, S., Schneeberger, K. & Ratsch, G. Optimal spliced alignments of short sequence reads. *Bioinformatics* **24**, i174-180, doi:10.1093/bioinformatics/btn300 (2008).

de Godoy, L. M. *et al.* Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome biology* **7**, R50, doi:10.1186/gb-2006-7-6-r50 (2006).

de Mercoyrol, L., Corda, Y., Job, C. & Job, D. Accuracy of wheat-germ RNA polymerase II. General enzymatic properties and effect of template conformational transition from right-handed B-DNA to left-handed Z-DNA. *European journal of biochemistry / FEBS* **206**, 49-58 (1992).

Denoeud, F. *et al.* Annotating genomes with massive-scale RNA sequencing. *Genome biology* **9**, R175, doi:10.1186/gb-2008-9-12-r175 (2008).

DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498, doi:10.1038/ng.806 (2011).

DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature genetics* **14**, 457-460, doi:10.1038/ng1296-457 (1996).

Deutsch, E. W. *et al.* A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **10**, 1150-1159, doi:10.1002/pmic.200900375 (2010).

DeVeale, B., van der Kooy, D. & Babak, T. Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. *PLoS genetics* **8**, e1002600, doi:10.1371/journal.pgen.1002600 (2012).

Di Giammartino, D. C., Nishida, K. & Manley, J. L. Mechanisms and consequences of alternative polyadenylation. *Molecular cell* **43**, 853-866, doi:10.1016/j.molcel.2011.08.017 (2011).

Diguistini, S. *et al.* De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome biology* **10**, R94, doi:10.1186/gb-2009-10-9-r94 (2009).

Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nature genetics* **39**, 1202-1207, doi:10.1038/ng2109 (2007).

Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108, doi:10.1038/nature11233 (2012).

Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome research* **17**, 1697-1706, doi:10.1101/gr.6435207 (2007).

Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research* **36**, e105, doi:10.1093/nar/gkn425 (2008).

Dolan, M. E. *et al.* Heritability and linkage analysis of sensitivity to cisplatin-induced cytotoxicity. *Cancer research* **64**, 4353-4356, doi:10.1158/0008-5472.CAN-04-0340 (2004).

Dorrell, R. G. & Howe, C. J. Functional remodeling of RNA processing in replacement chloroplasts by pathways retained from their predecessors. *Proceedings of the National Academy of Sciences of the United States of America*, doi:10.1073/pnas.1212270109 (2012).

Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 8817-8822, doi:10.1073/pnas.1133470100 (2003).

Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81, doi:10.1126/science.1181498 (2010).

Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

Eaves, H. L. & Gao, Y. MOM: maximum oligonucleotide mapping. *Bioinformatics* **25**, 969-970, doi:10.1093/bioinformatics/btp092 (2009).

Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138, doi:10.1126/science.1162986 (2009).

Elias, J. E., Haas, W., Faherty, B. K. & Gygi, S. P. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nature methods* **2**, 667-675, doi:10.1038/nmeth785 (2005).

Erie, D. A., Hajiseyedjavadi, O., Young, M. C. & von Hippel, P. H. Multiple RNA polymerase conformations and GreA: control of the fidelity of transcription. *Science* **262**, 867-873 (1993).

Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research* **8**, 186-194 (1998).

Fang, Y. *et al.* A complete sequence and transcriptomic analyses of date palm (Phoenix dactylifera L.) mitochondrial genome. *PloS one* **7**, e37164, doi:10.1371/journal.pone.0037164 (2012).

Farnham, P. J. Insights from genomic profiling of transcription factors. *Nature reviews. Genetics* **10**, 605-616, doi:10.1038/nrg2636 (2009).

Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic acids research* **34**, e22, doi:10.1093/nar/gnj023 (2006).

Feng, J. *et al.* GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-Seq data. *Bioinformatics*, doi:10.1093/bioinformatics/bts515 (2012).

Flicek, P. & Birney, E. Sense from sequence reads: methods for alignment and assembly. *Nature methods* **6**, S6-S12, doi:10.1038/nmeth.1376 (2009).

Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods* **7**, 461-465, doi:10.1038/nmeth.1459 (2010).

Fodor, S. P. *et al.* Multiplexed biochemical assays with biological chips. *Nature* **364**, 555-556, doi:10.1038/364555a0 (1993).

Freyer, R., Kiefer-Meyer, M. C. & Kossel, H. Occurrence of plastid RNA editing in all major lineages of land plants. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 6285-6290 (1997).

Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 1513-1518, doi:10.1073/pnas.1017351108 (2011).

Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537 (1999).

Gott, J. M. & Emeson, R. B. Functions and mechanisms of RNA editing. *Annual review of genetics* **34**, 499-531, doi:10.1146/annurev.genet.34.1.499 (2000).

Gott, J. M., Visomirski, L. M. & Hunter, J. L. Substitutional and insertional RNA editing of the cytochrome c oxidase subunit 1 mRNA of Physarum polycephalum. *The Journal of biological chemistry* **268**, 25483-25486 (1993).

Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**, 644-652, doi:10.1038/nbt.1883 (2011).

Grant, G. R. *et al.* Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* **27**, 2518-2528, doi:10.1093/bioinformatics/btr427 (2011).

Graveley, B. R. *et al.* The developmental transcriptome of Drosophila melanogaster. *Nature* **471**, 473-479, doi:10.1038/nature09715 (2011).

Greenberger, S. *et al.* Consistent levels of A-to-I RNA editing across individuals in coding sequences and non-conserved Alu repeats. *BMC genomics* **11**, 608, doi:10.1186/1471-2164-11-608 (2010).

Greeve, J., Altkemper, I., Dieterich, J. H., Greten, H. & Windler, E. Apolipoprotein B mRNA editing in 12 different mammalian species: hepatic expression is reflected in low concentrations of apoB-containing plasma lipoproteins. *Journal of lipid research* **34**, 1367-1383 (1993).

Gregg, C. *et al.* High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science* **329**, 643-648, doi:10.1126/science.1190830 (2010).

Grewe, F., Viehoever, P., Weisshaar, B. & Knoop, V. A trans-splicing group I intron and tRNA-hyperediting in the mitochondrial genome of the lycophyte Isoetes engelmannii. *Nucleic acids research* **37**, 5093-5104, doi:10.1093/nar/gkp532 (2009).

Grohmann, M. *et al.* Alternative splicing and extensive RNA editing of human TPH2 transcripts. *PloS one* **5**, e8956, doi:10.1371/journal.pone.0008956 (2010).

Gualberto, J. M., Lamattina, L., Bonnard, G., Weil, J. H. & Grienenberger, J. M. RNA editing in wheat mitochondria results in the conservation of protein sequences. *Nature* **341**, 660-662, doi:10.1038/341660a0 (1989).

Gualberto, J. M., Weil, J. H. & Grienenberger, J. M. Editing of the wheat coxIII transcript: evidence for twelve C to U and one U to C conversions and for sequence similarities around editing sites. *Nucleic acids research* **18**, 3771-3776 (1990).

Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology* **28**, 503-510, doi:10.1038/nbt.1633 (2010).

Halvardson, J., Zaghlool, A. & Feuk, L. Exome RNA sequencing reveals rare and novel alternative transcripts. *Nucleic acids research* **41**, e6, doi:10.1093/nar/gks816 (2013).

Harbers, M. & Carninci, P. Tag-based approaches for transcriptome research and genome annotation. *Nature methods* **2**, 495-502, doi:10.1038/nmeth768 (2005).

Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome biology* **10**, R32, doi:10.1186/gb-2009-10-3-r32 (2009).

Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome biology* **7 Suppl 1**, S4 1-9, doi:10.1186/gb-2006-7-s1-s4 (2006).

Heap, G. A. *et al.* Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Human molecular genetics* **19**, 122-134, doi:10.1093/hmg/ddp473 (2010).

Hendrickson, P. G. & Silliker, M. E. RNA editing in six mitochondrial ribosomal protein genes of Didymium iridis. *Current genetics* **56**, 203-213, doi:10.1007/s00294-010-0292-4 (2010).

Hernandez, D., Francois, P., Farinelli, L., Osteras, M. & Schrenzel, J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome research* **18**, 802-809, doi:10.1101/gr.072033.107 (2008).

Hersberger, M. & Innerarity, T. L. Two efficiency elements flanking the editing site of cytidine 6666 in the apolipoprotein B mRNA support mooring-dependent editing. *The Journal of biological chemistry* **273**, 9435-9442 (1998).

Hersberger, M., Patarroyo-White, S., Arnold, K. S. & Innerarity, T. L. Phylogenetic analysis of the apolipoprotein B mRNA-editing region. Evidence for a secondary structure between the mooring sequence and the 3' efficiency element. *The Journal of biological chemistry* **274**, 34590-34597 (1999).

Hiesel, R., Combettes, B. & Brennicke, A. Evidence for RNA editing in mitochondria of all major groups of land plants except the Bryophyta. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 629-633 (1994).

Hiesel, R., Wissinger, B., Schuster, W. & Brennicke, A. RNA editing in plant mitochondria. *Science* **246**, 1632-1634 (1989).

Higuchi, M. *et al.* Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* **406**, 78-81, doi:10.1038/35017558 (2000).

Hoch, B., Maier, R. M., Appel, K., Igloi, G. L. & Kossel, H. Editing of a chloroplast mRNA by creation of an initiation codon. *Nature* **353**, 178-180, doi:10.1038/353178a0 (1991).

Homer, N., Merriman, B. & Nelson, S. F. BFAST: an alignment tool for large scale genome resequencing. *PloS one* **4**, e7767, doi:10.1371/journal.pone.0007767 (2009).

Hoopengardner, B., Bhalla, T., Staber, C. & Reenan, R. Nervous system targets of RNA editing identified by comparative genomics. *Science* **301**, 832-836, doi:10.1126/science.1086763 (2003).

Hormozdiari, F., Alkan, C., Eichler, E. E. & Sahinalp, S. C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome research* **19**, 1270-1278, doi:10.1101/gr.088633.108 (2009).

Horton, T. L. & Landweber, L. F. Evolution of four types of RNA editing in myxomycetes. *RNA* **6**, 1339-1346 (2000).

Horton, T. L. & Landweber, L. F. Rewriting the information in DNA: RNA editing in kinetoplastids and myxomycetes. *Current opinion in microbiology* **5**, 620-626 (2002).

Hospattankar, A. V., Higuchi, K., Law, S. W., Meglin, N. & Brewer, H. B., Jr. Identification of a novel in-frame translational stop codon in human intestine apoB mRNA. *Biochemical and biophysical research communications* **148**, 279-285 (1987).

Hossain, M. S., Azimi, N. & Skiena, S. Crystallizing short-read assemblies around seeds. *BMC bioinformatics* **10 Suppl 1**, S16, doi:10.1186/1471-2105-10-S1-S16 (2009).

Hundley, H. A., Krauchuk, A. A. & Bass, B. L. C. elegans and H. sapiens mRNAs with edited 3' UTRs are present on polysomes. *RNA* **14**, 2050-2060, doi:10.1261/rna.1165008 (2008).

Idury, R. M. & Waterman, M. S. A new algorithm for DNA sequence assembly. *Journal of computational biology : a journal of computational molecular cell biology* **2**, 291-306 (1995).

Inada, M., Sasaki, T., Yukawa, M., Tsudzuki, T. & Sugiura, M. A systematic search for RNA editing sites in pea chloroplasts: an editing event causes diversification from the evolutionarily conserved amino acid sequence. *Plant & cell physiology* **45**, 1615-1622, doi:10.1093/pcp/pch191 (2004).

Innerarity, T. L. *et al.* Familial defective apolipoprotein B-100: a mutation of apolipoprotein B that causes hypercholesterolemia. *Journal of lipid research* **31**, 1337-1349 (1990).

International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789-796, doi:10.1038/nature02168 (2003).

International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-1320, doi:10.1038/nature04226 (2005).

Iyer, V. R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83-87 (1999).

Izban, M. G. & Luse, D. S. The RNA polymerase II ternary complex cleaves the nascent transcript in a 3'----5' direction in the presence of elongation factor SII. *Genes & development* **6**, 1342-1356 (1992).

Jackson, C. J. *et al.* Broad genomic and transcriptional analysis reveals a highly derived genome in dinoflagellate mitochondria. *BMC biology* **5**, 41, doi:10.1186/1741-7007-5-41 (2007).

Jeck, W. R. *et al.* Extending assembly of short DNA sequences to handle error. *Bioinformatics* **23**, 2942-2944, doi:10.1093/bioinformatics/btm451 (2007).

Jepson, J. E. & Reenan, R. A. RNA editing in regulating gene expression in the brain. *Biochimica et biophysica acta* **1779**, 459-470, doi:10.1016/j.bbagrm.2007.11.009 (2008).

Jepson, J. E., Savva, Y. A., Jay, K. A. & Reenan, R. A. Visualizing adenosine-to-inosine RNA editing in the Drosophila nervous system. *Nature methods* **9**, 189-194, doi:10.1038/nmeth.1827 (2012).

Jiang, H. & Wong, W. H. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **24**, 2395-2396, doi:10.1093/bioinformatics/btn429 (2008).

Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497-1502, doi:10.1126/science.1141319 (2007).

Ju, Y. S. *et al.* Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nature genetics* **43**, 745-752, doi:10.1038/ng.872 (2011).

Kane, J. P., Hardman, D. A. & Paulus, H. E. Heterogeneity of apolipoprotein B: isolation of a new species from human chylomicrons. *Proceedings of the National Academy of Sciences of the United States of America* **77**, 2465-2469 (1980).

Kao, W. C. & Song, Y. S. naiveBayesCall: an efficient model-based base-calling algorithm for high-throughput sequencing. *Journal of computational biology : a journal of computational molecular cell biology* **18**, 365-377, doi:10.1089/cmb.2010.0247 (2011).

Kao, W. C., Stevens, K. & Song, Y. S. BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing. *Genome research* **19**, 1884-1895, doi:10.1101/gr.095299.109 (2009).

Karijolich, J. & Yu, Y. T. Converting nonsense codons into sense codons by targeted pseudouridylation. *Nature* **474**, 395-398, doi:10.1038/nature10165 (2011).

Kashkina, E. *et al.* Template misalignment in multisubunit RNA polymerases and transcription fidelity. *Molecular cell* **24**, 257-266, doi:10.1016/j.molcel.2006.10.001 (2006).

Kawahara, Y. *et al.* Glutamate receptors: RNA editing and death of motor neurons. *Nature* **427**, 801, doi:10.1038/427801a (2004).

Kawahara, Y. *et al.* Frequency and fate of microRNA editing in human brain. *Nucleic acids research* **36**, 5270-5280, doi:10.1093/nar/gkn479 (2008).

Kawahara, Y. *et al.* Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* **315**, 1137-1140, doi:10.1126/science.1138050 (2007).

Keegan, L. P., Gallo, A. & O'Connell, M. A. The many roles of an RNA editor. *Nature reviews. Genetics* **2**, 869-878, doi:10.1038/35098584 (2001).

Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome research* **12**, 656-664, doi:10.1101/gr.229202. Article published online before March 2002 (2002).

Kent, W. J. *et al.* The human genome browser at UCSC. *Genome research* **12**, 996-1006, doi:10.1101/gr.229102. Article published online before print in May 2002 (2002).

Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103-107, doi:10.1038/nature09322 (2010).

Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64, doi:10.1038/nature06862 (2008).

Kim, J. B. *et al.* Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481-1484, doi:10.1126/science.1137325 (2007).

Kim, J. I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011-1015, doi:10.1038/nature08211 (2009).

Kim, S. R. *et al.* Rice OGR1 encodes a pentatricopeptide repeat-DYW protein and is essential for RNA editing in mitochondria. *The Plant journal : for cell and molecular biology* **59**, 738-749, doi:10.1111/j.1365-313X.2009.03909.x (2009).

Kim, Y. J. *et al.* ProbeMatch: rapid alignment of oligonucleotides to genome allowing both gaps and mismatches. *Bioinformatics* **25**, 1424-1425, doi:10.1093/bioinformatics/btp178 (2009).

Kiran, A. & Baranov, P. V. DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics* **26**, 1772-1776, doi:10.1093/bioinformatics/btq285 (2010).

Kircher, M., Stenzel, U. & Kelso, J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome biology* **10**, R83, doi:10.1186/gb-2009-10-8-r83 (2009).

Kireeva, M. L. *et al.* Transient reversal of RNA polymerase II active site closing controls fidelity of transcription elongation. *Molecular cell* **30**, 557-566, doi:10.1016/j.molcel.2008.04.017 (2008).

Kleinman, C. L., Adoue, V. & Majewski, J. RNA editing of protein sequences: A rare event in human transcriptomes. *RNA* **18**, 1586-1596, doi:10.1261/rna.033233.112 (2012).

Kleinman, C. L. & Majewski, J. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* **335**, 1302; author reply 1302, doi:10.1126/science.1209658 (2012).

Knoop, V. When you can't trust the DNA: RNA editing changes transcript sequences. *Cellular and molecular life sciences : CMLS* **68**, 567-586, doi:10.1007/s00018-010-0538-9 (2011).

Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283-2285, doi:10.1093/bioinformatics/btp373 (2009).

Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nature methods* **3**, 211-222, doi:10.1038/nmeth0306-211 (2006).

Korbel, J. O. *et al.* PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome biology* **10**, R23, doi:10.1186/gb-2009-10-2-r23 (2009).

Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420-426, doi:10.1126/science.1149504 (2007).

Kotera, E., Tasaka, M. & Shikanai, T. A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. *Nature* **433**, 326-330, doi:10.1038/nature03229 (2005).

Krawitz, P. *et al.* Microindel detection in short-read sequence data. *Bioinformatics* **26**, 722-729, doi:10.1093/bioinformatics/btq027 (2010).

Kugita, M., Yamamoto, Y., Fujikawa, T., Matsumoto, T. & Yoshinaga, K. RNA editing in hornwort chloroplasts makes more than half the genes functional. *Nucleic acids research* **31**, 2417-2423 (2003).

Kuhn, C. D. *et al.* Functional architecture of RNA polymerase I. *Cell* **131**, 1260-1272, doi:10.1016/j.cell.2007.10.051 (2007).

Landegren, U., Kaiser, R., Sanders, J. & Hood, L. A ligase-mediated gene detection technique. *Science* **241**, 1077-1080 (1988).

Langmead, B., Hansen, K. D. & Leek, J. T. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome biology* **11**, R83, doi:10.1186/gb-2010-11-8-r83 (2010).

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

Lau, N. C. *et al.* Characterization of the piRNA complex from rat testes. *Science* **313**, 363-367, doi:10.1126/science.1130164 (2006).

Laurencikiene, J., Kallman, A. M., Fong, N., Bentley, D. L. & Ohman, M. RNA editing and alternative splicing: the importance of co-transcriptional coordination. *EMBO reports* **7**, 303-307, doi:10.1038/sj.embor.7400621 (2006).

Le, S. Q. & Durbin, R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome research* **21**, 952-960, doi:10.1101/gr.113084.110 (2011).

Leamon, J. H. *et al.* A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* **24**, 3769-3777, doi:10.1002/elps.200305646 (2003).

Lee, R. D., Song, M. Y. & Lee, J. K. Large-scale profiling and identification of potential regulatory mechanisms for allelic gene expression in colorectal cancer cells. *Gene* **512**, 16-22, doi:10.1016/j.gene.2012.10.001 (2013).

Lee, S., Hormozdiari, F., Alkan, C. & Brudno, M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nature methods* **6**, 473-474, doi:10.1038/nmeth.f.256 (2009).

Legendre, P., Forstera, B., Juttner, R. & Meier, J. C. Glycine Receptors Caught between Genome and Proteome - Functional Implications of RNA Editing and Splicing. *Frontiers in molecular neuroscience* **2**, 23, doi:10.3389/neuro.02.023.2009 (2009).

Levanon, E. Y. *et al.* Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nature biotechnology* **22**, 1001-1005, doi:10.1038/nbt996 (2004).

Levin, J. Z. *et al.* Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome biology* **10**, R115, doi:10.1186/gb-2009-10-10-r115 (2009).

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493-500, doi:10.1093/bioinformatics/btp692 (2010).

Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18**, 1851-1858, doi:10.1101/gr.078212.108 (2008).

Li, J. B. *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**, 1210-1213, doi:10.1126/science.1170995 (2009).

Li, M., Nordborg, M. & Li, L. M. Adjust quality scores from alignment and improve sequencing accuracy. *Nucleic acids research* **32**, 5183-5191, doi:10.1093/nar/gkh850 (2004).

Li, M. *et al.* Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333**, 53-58, doi:10.1126/science.1207018 (2011).

Li, R. *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311-317, doi:10.1038/nature08696 (2010).

Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome research* **19**, 1124-1132, doi:10.1101/gr.088013.108 (2009).

Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713-714, doi:10.1093/bioinformatics/btn025 (2008).

Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967, doi:10.1093/bioinformatics/btp336 (2009).

Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20**, 265-272, doi:10.1101/gr.097261.109 (2010).

Liang, H. & Landweber, L. F. Hypothesis: RNA editing of microRNA target sites in humans? *RNA* **13**, 463-467, doi:10.1261/rna.296407 (2007).

Libby, R. T. & Gallant, J. A. The role of RNA polymerase in transcriptional fidelity. *Molecular microbiology* **5**, 999-1004 (1991).

Licatalosi, D. D. & Darnell, R. B. RNA processing and its regulation: global insights into biological networks. *Nature reviews. Genetics* **11**, 75-87, doi:10.1038/nrg2673 (2010).

Lin, H., Zhang, Z., Zhang, M. Q., Ma, B. & Li, M. ZOOM! Zillions of oligos mapped. *Bioinformatics* **24**, 2431-2437, doi:10.1093/bioinformatics/btn416 (2008).

Lin, S., Zhang, H., Spencer, D. F., Norman, J. E. & Gray, M. W. Widespread and extensive editing of mitochondrial mRNAS in dinoflagellates. *Journal of molecular biology* **320**, 727-739 (2002).

Lin, W., Piskol, R., Tan, M. H. & Li, J. B. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* **335**, 1302; author reply 1302, doi:10.1126/science.1210624 (2012).

Linsley, P. S. *et al.* CTLA-4 is a second receptor for the B cell activation antigen B7. *The Journal of experimental medicine* **174**, 561-569 (1991).

Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523-536, doi:10.1016/j.cell.2008.03.029 (2008).

Locke, D. P. *et al.* Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529-533, doi:10.1038/nature09687 (2011).

Lomeli, H. *et al.* Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. *Science* **266**, 1709-1713 (1994).

Lu, T. *et al.* Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome research* **20**, 1238-1249, doi:10.1101/gr.106120.110 (2010).

Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research* **21**, 936-939, doi:10.1101/gr.111120.110 (2011).

Luo, C., Tsementzi, D., Kyrpides, N., Read, T. & Konstantinidis, K. T. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PloS one* **7**, e30087, doi:10.1371/journal.pone.0030087 (2012).

Lutz, C. S. Alternative polyadenylation: a twist on mRNA 3' end formation. *ACS chemical biology* **3**, 609-617, doi:10.1021/cb800138w (2008).

Maas, S., Patt, S., Schrey, M. & Rich, A. Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. *Proceedings of the National Academy of Sciences of*

the United States of America **98**, 14687-14692, doi:10.1073/pnas.251531398 (2001).

Mahendran, R., Spottswood, M. R. & Miller, D. L. RNA editing by cytidine insertion in mitochondria of Physarum polycephalum. *Nature* **349**, 434-438, doi:10.1038/349434a0 (1991).

Maher, C. A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97-101, doi:10.1038/nature07638 (2009).

Malek, O., Lattig, K., Hiesel, R., Brennicke, A. & Knoop, V. RNA editing in bryophytes and a molecular phylogeny of land plants. *The EMBO journal* **15**, 1403-1411 (1996).

Malone, C. D. & Hannon, G. J. Small RNAs as guardians of the genome. *Cell* **136**, 656-668, doi:10.1016/j.cell.2009.01.045 (2009).

Maniatis, T. & Tasic, B. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**, 236-243, doi:10.1038/418236a (2002).

Mardis, E. R. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics* **9**, 387-402, doi:10.1146/annurev.genom.9.081307.164359 (2008).

Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380, doi:10.1038/nature03959 (2005).

Martin, G. & Keller, W. RNA-specific ribonucleotidyl transferases. *RNA* **13**, 1834-1849, doi:10.1261/rna.652807 (2007).

Martin, J. *et al.* Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC genomics* **11**, 663, doi:10.1186/1471-2164-11-663 (2010).

Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nature reviews. Genetics* **12**, 671-682, doi:10.1038/nrg3068 (2011).

Martinez, H. D. *et al.* RNA editing of androgen receptor gene transcripts in prostate cancer cells. *The Journal of biological chemistry* **283**, 29938-29949, doi:10.1074/jbc.M800534200 (2008).

Matise, T. C. *et al.* A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. *American journal of human genetics* **73**, 271-284 (2003).

McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome research* **19**, 1527-1541, doi:10.1101/gr.091868.109 (2009).

Meacham, F. *et al.* Identification and correction of systematic error in high-throughput sequence data. *BMC bioinformatics* **12**, 451, doi:10.1186/1471-2105-12-451 (2011).

Mehta, A., Kinter, M. T., Sherman, N. E. & Driscoll, D. M. Molecular cloning of apobec-1 complementation factor, a novel RNA-binding protein involved in the editing of apolipoprotein B mRNA. *Molecular and cellular biology* **20**, 1846-1854 (2000).

Meier, J. C. *et al.* RNA editing produces glycine receptor alpha3(P185L), resulting in high agonist potency. *Nature neuroscience* **8**, 736-744, doi:10.1038/nn1467 (2005).

Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-770, doi:10.1038/nature07107 (2008).

Mercer, T. R. *et al.* Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nature biotechnology* **30**, 99-104, doi:10.1038/nbt.2024 (2012).

Metzker, M. L. Emerging technologies in DNA sequencing. *Genome research* **15**, 1767-1776, doi:10.1101/gr.3770505 (2005).

Metzker, M. L. Sequencing in real time. *Nature biotechnology* **27**, 150-151, doi:10.1038/nbt0209-150 (2009).

Metzker, M. L. Sequencing technologies - the next generation. *Nature reviews. Genetics* **11**, 31-46, doi:10.1038/nrg2626 (2010).

Michalski, A., Cox, J. & Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *Journal of proteome research* **10**, 1785-1793, doi:10.1021/pr101060v (2011).

Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560, doi:10.1038/nature06008 (2007).

Miller, J. R. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818-2824, doi:10.1093/bioinformatics/btn548 (2008).

Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59-65, doi:10.1038/nature09708 (2011).

Minoche, A. E., Dohm, J. C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome biology* **12**, R112, doi:10.1186/gb-2011-12-11-r112 (2011).

Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-777, doi:10.1038/nature08903 (2010).

Moore, M. J. & Proudfoot, N. J. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* **136**, 688-700, doi:10.1016/j.cell.2009.02.001 (2009).

Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743-747, doi:10.1038/nature02797 (2004).

Morse, D. P. & Bass, B. L. Long RNA hairpins that contain inosine are present in Caenorhabditis elegans poly(A)+ RNA. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 6048-6053 (1999).

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621-628, doi:10.1038/nmeth.1226 (2008).

Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349, doi:10.1126/science.1158441 (2008).

Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic acids research* **39**, e90, doi:10.1093/nar/gkr344 (2011).

Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature genetics* **42**, 790-793, doi:10.1038/ng.646 (2010).

Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics* **42**, 30-35, doi:10.1038/ng.499 (2010).

Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-276, doi:10.1038/nature08250 (2009).

Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics* **12**, 443-451, doi:10.1038/nrg2986 (2011).

Nishikura, K. Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nature reviews. Molecular cell biology* **7**, 919-931, doi:10.1038/nrm2061 (2006).

Nishikura, K. Functions and regulation of RNA editing by ADAR deaminases. *Annual review of biochemistry* **79**, 321-349, doi:10.1146/annurev-biochem-060208-105251 (2010).

Novo, F. J., Kruszewski, A., MacDermot, K. D., Goldspink, G. & Gorecki, D. C. Editing of human alpha-galactosidase RNA resulting in a pyrimidine to purine conversion. *Nucleic acids research* **23**, 2636-2640 (1995).

O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature genetics* **43**, 585-589, doi:10.1038/ng.835 (2011).

Okoniewski, M. J. & Miller, C. J. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC bioinformatics* **7**, 276, doi:10.1186/1471-2105-7-276 (2006).

Okuda, K. *et al.* The pentatricopeptide repeat protein OTP82 is required for RNA editing of plastid ndhB and ndhG transcripts. *The Plant journal : for cell and molecular biology* **61**, 339-349, doi:10.1111/j.1365-313X.2009.04059.x (2010).

Orlandi, C., Barbon, A. & Barlati, S. Activity regulation of adenosine deaminases acting on RNA (ADARs). *Molecular neurobiology* **45**, 61-75, doi:10.1007/s12035-011-8220-2 (2012).

Ozsolak, F. *et al.* Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143**, 1018-1029, doi:10.1016/j.cell.2010.11.020 (2010).

Ozsolak, F. & Milos, P. M. Transcriptome profiling using single-molecule direct RNA sequencing. *Methods Mol Biol* **733**, 51-61, doi:10.1007/978-1-61779-089-8_4 (2011).

Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461**, 814-818, doi:10.1038/nature08390 (2009).

Palladino, M. J., Keegan, L. P., O'Connell, M. A. & Reenan, R. A. A-to-I pre-mRNA editing in Drosophila is primarily involved in adult nervous system function and integrity. *Cell* **102**, 437-449 (2000).

Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics* **40**, 1413-1415, doi:10.1038/ng.259 (2008).

Park, E., Williams, B., Wold, B. J. & Mortazavi, A. RNA editing in the human ENCODE RNA-Seq data. *Genome research* **22**, 1626-1633, doi:10.1101/gr.134957.111 (2012).

Paschen, W., Hedreen, J. C. & Ross, C. A. RNA editing of the glutamate receptor subunits GluR2 and GluR6 in human brain tissue. *Journal of neurochemistry* **63**, 1596-1602 (1994).

Peng, Z. *et al.* Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature biotechnology* **30**, 253-260, doi:10.1038/nbt.2122 (2012).

Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567, doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2 (1999).

Peters, P. J., Neefjes, J. J., Oorschot, V., Ploegh, H. L. & Geuze, H. J. Segregation of MHC class II molecules from MHC class I molecules in the Golgi complex for transport to lysosomal compartments. *Nature* **349**, 669-676, doi:10.1038/349669a0 (1991).

Pevzner, P. A., Borodovsky, M. & Mironov, A. A. Linguistics of nucleotide sequences. II: Stationary words in genetic texts and the zonal structure of DNA. *Journal of biomolecular structure & dynamics* **6**, 1027-1038 (1989).

Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 9748-9753, doi:10.1073/pnas.171285098 (2001).

Phreaner, C. G., Williams, M. A. & Mulligan, R. M. Incomplete editing of rps12 transcripts results in the synthesis of polymorphic polypeptides in plant mitochondria. *The Plant cell* **8**, 107-117, doi:10.1105/tpc.8.1.107 (1996).

Pickrell, J. K., Gilad, Y. & Pritchard, J. K. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* **335**, 1302; author reply 1302, doi:10.1126/science.1210484 (2012).

Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS genetics* **6**, e1001236, doi:10.1371/journal.pgen.1001236 (2010).

Polson, A. G. & Bass, B. L. Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. *The EMBO journal* **13**, 5701-5711 (1994).

Pomerantz, R. T., Temiakov, D., Anikin, M., Vassylyev, D. G. & McAllister, W. T. A mechanism of nucleotide misincorporation during transcription due to template-strand misalignment. *Molecular cell* **24**, 245-255, doi:10.1016/j.molcel.2006.08.014 (2006).

Powell, L. M. *et al.* A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* **50**, 831-840 (1987).

Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic acids research* **40**, D130-135, doi:10.1093/nar/gkr1079 (2012).

Pullirsch, D. & Jantsch, M. F. Proteome diversification by adenosine to inosine RNA editing. *RNA biology* **7**, 205-212 (2010).

Qi, J., Zhao, F., Buboltz, A. & Schuster, S. C. inGAP: an integrated next-generation genome analysis pipeline. *Bioinformatics* **26**, 127-129, doi:10.1093/bioinformatics/btp615 (2010).

Qian, L., Quadros, E. V., Regec, A., Zittoun, J. & Rothenberg, S. P. Congenital transcobalamin II deficiency due to errors in RNA editing. *Blood cells, molecules & diseases* **28**, 134-142; discussion 143-135 (2002).

Qu, W., Hashimoto, S. & Morishita, S. Efficient frequency-based de novo short-read clustering for error trimming in next-generation sequencing. *Genome research* **19**, 1309-1315, doi:10.1101/gr.089151.108 (2009).

Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics* **13**, 341, doi:10.1186/1471-2164-13-341 (2012).

Quinlan, A. R., Stewart, D. A., Stromberg, M. P. & Marth, G. T. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nature methods* **5**, 179-181, doi:10.1038/nmeth.1172 (2008).

Ramaswami, G. *et al.* Accurate identification of human Alu and non-Alu RNA editing sites. *Nature methods*, doi:10.1038/nmeth.1982 (2012).

Ramaswami, G. *et al.* Identifying RNA editing sites using RNA sequencing data alone. *Nature methods* **10**, 128-132, doi:10.1038/nmeth.2330 (2013).

Reid, J. G. *et al.* Mouse let-7 miRNA populations exhibit RNA editing that is constrained in the 5'-seed/ cleavage/anchor regions and stabilize predicted mmu-let-7a:mRNA duplexes. *Genome research* **18**, 1571-1581, doi:10.1101/gr.078246.108 (2008).

Robbins, J. C., Heller, W. P. & Hanson, M. R. A comparative genomics approach identifies a PPR-DYW protein that is essential for C-to-U editing of the Arabidopsis chloroplast accD transcript. *RNA* **15**, 1142-1153, doi:10.1261/rna.1533909 (2009).

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome biology* **12**, R22, doi:10.1186/gb-2011-12-3-r22 (2011).

Robertson, G. *et al.* De novo assembly and analysis of RNA-Seq data. *Nature methods* **7**, 909-912, doi:10.1038/nmeth.1517 (2010).

Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

Rodriguez, J., Menet, J. S. & Rosbash, M. Nascent-seq indicates widespread cotranscriptional RNA editing in Drosophila. *Molecular cell* **47**, 27-37, doi:10.1016/j.molcel.2012.05.002 (2012).

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. & Nyren, P. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry* **242**, 84-89, doi:10.1006/abio.1996.0432 (1996).

Ronaghi, M., Uhlen, M. & Nyren, P. A sequencing method based on real-time pyrophosphate. *Science* **281**, 363, 365 (1998).

Rosenberg, B. R., Hamilton, C. E., Mwangi, M. M., Dewell, S. & Papavasiliou, F. N. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. *Nature structural & molecular biology* **18**, 230-236, doi:10.1038/nsmb.1975 (2011).

Rosenberger, R. F. & Hilton, J. The frequency of transcriptional and translational errors at nonsense codons in the lacZ gene of Escherichia coli. *Molecular & general genetics : MGG* **191**, 207-212 (1983).

Rosenbloom, K. R. *et al.* ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic acids research* **38**, D620-625, doi:10.1093/nar/gkp961 (2010).

Rosenthal, J. J. & Bezanilla, F. Extensive editing of mRNAs for the squid delayed rectifier K+ channel regulates subunit tetramerization. *Neuron* **34**, 743-757 (2002).

Ross, R. Cell biology of atherosclerosis. *Annual review of physiology* **57**, 791-804, doi:10.1146/annurev.ph.57.030195.004043 (1995).

Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348-352, doi:10.1038/nature10242 (2011).

Rothberg, J. M. & Leamon, J. H. The development and impact of 454 sequencing. *Nature biotechnology* **26**, 1117-1124, doi:10.1038/nbt1485 (2008).

Rottman, F. M., Bokar, J. A., Narayan, P., Shambaugh, M. E. & Ludwiczak, R. N6-adenosine methylation in mRNA: substrate specificity and enzyme complexity. *Biochimie* **76**, 1109-1114 (1994).

Rueter, S. M., Burns, C. M., Coode, S. A., Mookherjee, P. & Emeson, R. B. Glutamate receptor RNA editing in vitro by enzymatic conversion of adenosine to inosine. *Science* **267**, 1491-1494 (1995).

Rueter, S. M., Dawson, T. R. & Emeson, R. B. Regulation of alternative splicing by RNA editing. *Nature* **399**, 75-80, doi:10.1038/19992 (1999).

Rumble, S. M. *et al.* SHRiMP: accurate mapping of short color-space reads. *PLoS computational biology* **5**, e1000386, doi:10.1371/journal.pcbi.1000386 (2009).

Sakurai, M., Yano, T., Kawabata, H., Ueda, H. & Suzuki, T. Inosine cyanoethylation identifies A-to-I RNA editing sites in the human transcriptome. *Nature chemical biology* **6**, 733-740, doi:10.1038/nchembio.434 (2010).

Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A. & Burge, C. B. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* **320**, 1643-1647, doi:10.1126/science.1155390 (2008).

Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463-5467 (1977).

Sawada, J. *et al.* Effects of antidepressants on GluR2 Q/R site-RNA editing in modified HeLa cell line. *Neuroscience research* **64**, 251-258, doi:10.1016/j.neures.2009.03.009 (2009).

Schaefer, M., Pollex, T., Hanna, K. & Lyko, F. RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic acids research* **37**, e12, doi:10.1093/nar/gkn954 (2009).

Schatz, M. C. CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics* **25**, 1363-1369, doi:10.1093/bioinformatics/btp236 (2009).

Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470 (1995).

Schoft, V. K., Schopoff, S. & Jantsch, M. F. Regulation of glutamate receptor B pre-mRNA splicing by RNA editing. *Nucleic acids research* **35**, 3723-3732, doi:10.1093/nar/gkm314 (2007).

Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887-898, doi:10.1016/j.cell.2008.02.022 (2008).

Schrider, D. R., Gout, J. F. & Hahn, M. W. Very few RNA and DNA sequence differences in the human transcriptome. *PloS one* **6**, e25842, doi:10.1371/journal.pone.0025842 (2011).

Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-Seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086-1092, doi:10.1093/bioinformatics/bts094 (2012).

Schuster, S. C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943-947, doi:10.1038/nature08795 (2010).

Schuster, W., Hiesel, R., Wissinger, B. & Brennicke, A. RNA editing in the cytochrome b locus of the higher plant Oenothera berteriana includes a U-to-C transition. *Molecular and cellular biology* **10**, 2428-2431 (1990).

Seeburg, P. H. & Hartner, J. Regulation of ion channel/neurotransmitter receptor function by RNA editing. *Current opinion in neurobiology* **13**, 279-283 (2003).

Shah, R. R. *et al.* Sequence requirements for the editing of apolipoprotein B mRNA. *The Journal of biological chemistry* **266**, 16301-16304 (1991).

Sharma, P. M., Bowman, M., Madden, S. L., Rauscher, F. J., 3rd & Sukumar, S. RNA editing in the Wilms' tumor susceptibility gene, WT1. *Genes & development* **8**, 720-731 (1994).

Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature biotechnology* **26**, 1135-1145, doi:10.1038/nbt1486 (2008).

Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732, doi:10.1126/science.1117389 (2005).

Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29**, 308-311 (2001).

Silberberg, G., Lundin, D., Navon, R. & Ohman, M. Deregulation of the A-to-I RNA editing mechanism in psychiatric disorders. *Human molecular genetics* **21**, 311-321, doi:10.1093/hmg/ddr461 (2012).

Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome research* **19**, 1117-1123, doi:10.1101/gr.089532.108 (2009).

Skuse, G. R., Cappione, A. J., Sowden, M., Metheny, L. J. & Smith, H. C. The neurofibromatosis type I messenger RNA undergoes base-modification RNA editing. *Nucleic acids research* **24**, 478-485 (1996).

Smit, A. F. The origin of interspersed repeats in the human genome. *Current opinion in genetics & development* **6**, 743-748 (1996).

Smith, A. D., Xuan, Z. & Zhang, M. Q. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC bioinformatics* **9**, 128, doi:10.1186/1471-2105-9-128 (2008).

Smith, D. R. *et al.* Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome research* **18**, 1638-1642, doi:10.1101/gr.077776.108 (2008).

Smith, H. C., Bennett, R. P., Kizilyer, A., McDougall, W. M. & Prohaska, K. M. Functions and regulation of the APOBEC family of proteins. *Seminars in cell & developmental biology* **23**, 258-268, doi:10.1016/j.semcdb.2011.10.004 (2012).

Smith, H. C., Gott, J. M. & Hanson, M. R. A guide to RNA editing. *RNA* **3**, 1105-1123 (1997).

Sodek, K. L. *et al.* Identification of pathways associated with invasive behavior by ovarian cancer cells using multidimensional protein identification technology (MudPIT). *Molecular bioSystems* **4**, 762-773, doi:10.1039/b717542f (2008).

Sommer, B., Kohler, M., Sprengel, R. & Seeburg, P. H. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* **67**, 11-19 (1991).

Song, C. X., Yi, C. & He, C. Mapping recently identified nucleotide variants in the genome and transcriptome. *Nature biotechnology* **30**, 1107-1116, doi:10.1038/nbt.2398 (2012).

Sper-Whitis, G. L., Moody, J. L. & Vaughn, J. C. Universality of mitochondrial RNA editing in cytochrome-c oxidase subunit I (coxI) among the land plants. *Biochimica et biophysica acta* **1307**, 301-308 (1996).

Sper-Whitis, G. L., Russell, A. L. & Vaughn, J. C. Mitochondrial RNA editing of cytochrome c oxidase subunit II (coxII) in the primitive vascular plant Psilotum nudum. *Biochimica et biophysica acta* **1218**, 218-220 (1994).

Steinhauser, S., Beckert, S., Capesius, I., Malek, O. & Knoop, V. Plant mitochondrial RNA editing. *Journal of molecular evolution* **48**, 303-312 (1999).

Stern, A. S. *et al.* Purification to homogeneity and partial characterization of interleukin 2 from a human T-cell leukemia. *Proceedings of the National Academy of Sciences of the United States of America* **81**, 871-875 (1984).

Sultan, M. *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956-960, doi:10.1126/science.1160342 (2008).

Surget-Groba, Y. & Montoya-Burgos, J. I. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome research* **20**, 1432-1440, doi:10.1101/gr.103846.109 (2010).

Sutton, C. A., Zoubenko, O. V., Hanson, M. R. & Maliga, P. A plant mitochondrial sequence transcribed in transgenic tobacco chloroplasts is not edited. *Molecular and cellular biology* **15**, 1377-1381 (1995).

Sydow, J. F. *et al.* Structural basis of transcription: mismatch-specific fidelity mechanisms and paused RNA polymerase II with frayed RNA. *Molecular cell* **34**, 710-721, doi:10.1016/j.molcel.2009.06.002 (2009).

Sydow, J. F. & Cramer, P. RNA polymerase fidelity and transcriptional proofreading. *Current opinion in structural biology* **19**, 732-739, doi:10.1016/j.sbi.2009.10.009 (2009).

Taft, R. J. *et al.* Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nature structural & molecular biology* **17**, 1030-1034, doi:10.1038/nsmb.1841 (2010).

Tan, B. Z., Huang, H., Lam, R. & Soong, T. W. Dynamic regulation of RNA editing of ion channels and receptors in the mammalian nervous system. *Molecular brain* **2**, 13, doi:10.1186/1756-6606-2-13 (2009).

Temiakov, D. *et al.* Structural basis for substrate selection by t7 RNA polymerase. *Cell* **116**, 381-391 (2004).

Temple, G. *et al.* The completion of the Mammalian Gene Collection (MGC). *Genome research* **19**, 2324-2333, doi:10.1101/gr.095976.109 (2009).

Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-69, doi:10.1126/science.1219240 (2012).

Thomas, M. J., Platas, A. A. & Hawley, D. K. Transcriptional fidelity and proofreading by RNA polymerase II. *Cell* **93**, 627-637 (1998).

Thomas, S. M., Lamb, R. A. & Paterson, R. G. Two mRNAs that differ by two nontemplated nucleotides encode the amino coterminal proteins P and V of the paramyxovirus SV5. *Cell* **54**, 891-902 (1988).

Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, doi:10.1093/bib/bbs017 (2012).

Tillich, M. *et al.* Editing of plastid RNA in Arabidopsis thaliana ecotypes. *The Plant journal : for cell and molecular biology* **43**, 708-715, doi:10.1111/j.1365-313X.2005.02484.x (2005).

Tonkin, L. A. *et al.* RNA editing by ADARs is important for normal behavior in Caenorhabditis elegans. *The EMBO journal* **21**, 6025-6035 (2002).

Toung, J. M., Morley, M., Li, M. & Cheung, V. G. RNA-sequence analysis of human B-cells. *Genome research* **21**, 991-998, doi:10.1101/gr.116335.110 (2011).

Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111, doi:10.1093/bioinformatics/btp120 (2009).

Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562-578, doi:10.1038/nprot.2012.016 (2012).

Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511-515, doi:10.1038/nbt.1621 (2010).

Tsudzuki, T., Wakasugi, T. & Sugiura, M. Comparative analysis of RNA editing sites in higher plant chloroplasts. *Journal of molecular evolution* **53**, 327-332, doi:10.1007/s002390010222 (2001).

Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484, doi:10.1038/nature07540 (2009).

Turner, T. L., Bourne, E. C., Von Wettberg, E. J., Hu, T. T. & Nuzhdin, S. V. Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. *Nature genetics* **42**, 260-263, doi:10.1038/ng.515 (2010).

Twine, N. A., Janitz, K., Wilkins, M. R. & Janitz, M. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. *PloS one* **6**, e16266, doi:10.1371/journal.pone.0016266 (2011).

Underwood, J. G. *et al.* FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature methods* **7**, 995-1001, doi:10.1038/nmeth.1529 (2010).

van Leeuwen, F. W. *et al.* Frameshift mutants of beta amyloid precursor protein and ubiquitin-B in Alzheimer's and Down patients. *Science* **279**, 242-247 (1998).

Vassylyev, D. G. *et al.* Structural basis for substrate loading in bacterial RNA polymerase. *Nature* **448**, 163-168, doi:10.1038/nature05931 (2007).

Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484-487 (1995).

Vidal, S., Curran, J. & Kolakofsky, D. Editing of the Sendai virus P/C mRNA by G insertion occurs during mRNA synthesis via a virus-encoded activity. *Journal of virology* **64**, 239-246 (1990a).

Vidal, S., Curran, J. & Kolakofsky, D. A stuttering model for paramyxovirus P mRNA editing. *The EMBO journal* **9**, 2017-2022 (1990b).

Wahlstedt, H., Daniel, C., Enstero, M. & Ohman, M. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome research* **19**, 978-986, doi:10.1101/gr.089409.108 (2009).

Wang, D. *et al.* Structural basis of transcription: backtracked RNA polymerase II at 3.4 angstrom resolution. *Science* **324**, 1203-1206, doi:10.1126/science.1168729 (2009).

Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476, doi:10.1038/nature07509 (2008).

Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60-65, doi:10.1038/nature07484 (2008).

Wang, K. *et al.* MapSplice: accurate mapping of RNA-Seq reads for splice junction discovery. *Nucleic acids research* **38**, e178, doi:10.1093/nar/gkq622 (2010).

Wang, L., Feng, Z., Wang, X., Wang, X. & Zhang, X. DEGseq: an R package for identifying differentially expressed genes from RNA-Seq data. *Bioinformatics* **26**, 136-138, doi:10.1093/bioinformatics/btp612 (2010).

Wang, Q., Khillan, J., Gadue, P. & Nishikura, K. Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis. *Science* **290**, 1765-1768 (2000).

Wang, W., Wu, Y. & Messing, J. The Mitochondrial Genome of an Aquatic Plant, Spirodela polyrhiza. *PloS one* **7**, e46747, doi:10.1371/journal.pone.0046747 (2012).

Warf, M. B., Shepherd, B. A., Johnson, W. E. & Bass, B. L. Effects of ADARs on small RNA processing pathways in C. elegans. *Genome research* **22**, 1488-1498, doi:10.1101/gr.134841.111 (2012).

Warren, R. L., Sutton, G. G., Jones, S. J. & Holt, R. A. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23**, 500-501, doi:10.1093/bioinformatics/btl629 (2007).

Wedekind, J. E., Dance, G. S., Sowden, M. P. & Smith, H. C. Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. *Trends in genetics : TIG* **19**, 207-216, doi:10.1016/S0168-9525(03)00054-4 (2003).

Wederell, E. D. *et al.* Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic acids research* **36**, 4549-4564, doi:10.1093/nar/gkn382 (2008).

Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876, doi:10.1038/nature06884 (2008).

Wilhelm, B. T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239-1243, doi:10.1038/nature07002 (2008).

Wolf, P. G., Rowe, C. A. & Hasebe, M. High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern Adiantum capillus-veneris. *Gene* **339**, 89-97, doi:10.1016/j.gene.2004.06.018 (2004).

Wommack, K. E., Bhavsar, J. & Ravel, J. Metagenomics: read length matters. *Applied and environmental microbiology* **74**, 1453-1463, doi:10.1128/AEM.02181-07 (2008).

Wu, D., Lamm, A. T. & Fire, A. Z. Competition between ADAR and RNAi pathways for an extensive class of RNA targets. *Nature structural & molecular biology* **18**, 1094-1101, doi:10.1038/nsmb.2129 (2011).

Wu, H., Irizarry, R. A. & Bravo, H. C. Intensity normalization improves color calling in SOLiD sequencing. *Nature methods* **7**, 336-337, doi:10.1038/nmeth0510-336 (2010).

Wu, H., Wang, C. & Wu, Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-Seq data. *Biostatistics*, doi:10.1093/biostatistics/kxs033 (2012).

Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873-881, doi:10.1093/bioinformatics/btq057 (2010).

Wulff, B. E. & Nishikura, K. Substitutional A-to-I RNA editing. *Wiley interdisciplinary reviews. RNA* **1**, 90-101, doi:10.1002/wrna.10 (2010).

Xia, Q. *et al.* Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). *Science* **326**, 433-436, doi:10.1126/science.1176620 (2009).

Yamanaka, S., Poksay, K. S., Arnold, K. S. & Innerarity, T. L. A novel translational repressor mRNA is edited extensively in livers containing tumors caused by the transgene expression of the apoB mRNA-editing enzyme. *Genes & development* **11**, 321-333 (1997).

Yang, W. *et al.* Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nature structural & molecular biology* **13**, 13-21, doi:10.1038/nsmb1041 (2006).

Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871, doi:10.1093/bioinformatics/btp394 (2009).

Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75-78, doi:10.1126/science.1190371 (2010).

Yoshinaga, K., Iinuma, H., Masuzawa, T. & Uedal, K. Extensive RNA editing of U to C in addition to C to U substitution in the rbcL transcripts of hornwort chloroplasts and the origin of RNA editing in green plants. *Nucleic acids research* **24**, 1008-1014 (1996).

Yu, Q. B., Jiang, Y., Chong, K. & Yang, Z. N. AtECB2, a pentatricopeptide repeat protein, is required for chloroplast transcript accD RNA editing and early chloroplast biogenesis in Arabidopsis thaliana. *The Plant journal : for cell and molecular biology* **59**, 1011-1023, doi:10.1111/j.1365-313X.2009.03930.x (2009).

Zauner, S., Greilinger, D., Laatsch, T., Kowallik, K. V. & Maier, U. G. Substitutional editing of transcripts from genes of cyanobacterial origin in the dinoflagellate Ceratium horridum. *FEBS letters* **577**, 535-538, doi:10.1016/j.febslet.2004.10.060 (2004).

Zehrmann, A., Verbitskiy, D., van der Merwe, J. A., Brennicke, A. & Takenaka, M. A DYW domain-containing pentatricopeptide repeat protein is required for RNA editing at multiple sites in mitochondria of Arabidopsis thaliana. *The Plant cell* **21**, 558-567, doi:10.1105/tpc.108.064535 (2009).

Zenkin, N., Yuzenkova, Y. & Severinov, K. Transcript-assisted transcriptional proofreading. *Science* **313**, 518-520, doi:10.1126/science.1127422 (2006).

Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-829, doi:10.1101/gr.074492.107 (2008).

Zhang, W. *et al.* A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PloS one* **6**, e17915, doi:10.1371/journal.pone.0017915 (2011).

Zhou, W. *et al.* The Arabidopsis gene YS1 encoding a DYW protein is required for editing of rpoB transcripts and the rapid development of chloroplasts during early growth. *The Plant journal : for cell and molecular biology* **58**, 82-96, doi:10.1111/j.1365-313X.2008.03766.x (2009).

Zhu, H. *et al.* Quantitative analysis of focused a-to-I RNA editing sites by ultra-high-throughput sequencing in psychiatric disorders. *PloS one* **7**, e43227, doi:10.1371/journal.pone.0043227 (2012).